AD_____

Award Number: DAMD17-00-1-0197

TITLE: Computer-Aided Diagnosis of Digital Mammograms

PRINCIPAL INVESTIGATOR: Yulei Jiang, Ph.D.

CONTRACTING ORGANIZATION: The University of Chicago
Chicago, Illinois 60637

REPORT DATE: June 2003

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	June 2003	Annual Summary (1 Jun 2002 - 31 May 2003)

4. TITLE AND SUBTITLE
Computer-Aided Diagnosis of Digital Mammograms

5. FUNDING NUMBERS
DAMD17-00-1-0197

6. AUTHOR(S)
Yulei Jiang, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The University of Chicago
Chicago, Illinois 60637

E-Mail: y-jiang@uchicago.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

The long-term goal of our research is to develop computer-aided diagnosis (CAD) techniques to improve the detection and diagnosis of breast cancer. We have developed a computer technique that can classify breast calcifications in mammograms accurately, and this technique as a diagnostic aid has been shown to be able to improve radiologists' diagnostic accuracy. We have determined that Breast Imaging Report and Data System (BI-RADS) lesion descriptions provided by radiologists can be used as supplemental data to computer-extracted image features to improve the performance of computer classification of malignant and benign breast lesions. We have also found that our classification technique developed on screen-film mammograms, can achieve equality high performance on full-filed digital mammograms. This high performance is little affected by variability in the way in which radiologists indicate the general location of calcification to the computer, which is designed as a means for the radiologist to query the computer aid. These results suggest that the computer technique has the potential to become a clinically useful and viable tool for diagnostic mammography.

14. SUBJECT TERMS
Computer-aided diagnosis (CAD), full-field digital mammography, BI-RADS, ROC analysis, artificial neural network (ANN)

15. NUMBER OF PAGES
54

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

TABLE OF CONTENTS

INTRODUCTION

The long-term goal of our research is to develop computer-aided diagnosis (CAD) techniques to improve the detection and diagnosis of breast cancer. The hypothesis to be tested in the present project is that radiologists' ability to differentiate malignant from benign breast lesions can be improved by integrating radiologists' perceptual expertise in the interpretation of mammograms with the advantages of automated computer classification. This project has three SOW Tasks:

Task 1. To combine radiologist-extracted Breast Imaging Reporting and Data System (BI-RADS) features with image features extracted by a computer to classify malignant and benign clustered microcalcifications in mammograms.

Task 2. To optimally combine radiologists' diagnosis with the result of computer classification.

Task 3. To optimize computer classification for full-field digital mammograms.

BODY

1. Combination of BI-RADS features and computer-extracted image features for computer classification of malignant and benign breast lesions

This particular work is within the scope of SOW Task 1. We have continued from what was reported in the last report to work on the investigation of BI-RADS lesion descriptors for computer-aided diagnosis of malignant and benign breast lesions in mammograms. We have used the database assembled earlier and have now completed the study of three expert mammography specialists reading the mammograms. In the experiment, the radiologists read the standard view and magnification view screen-film mammograms that contain suspicious lesions. Original film mammograms were used in the study to simulate clinical practice. The radiologists then provided standard BI-RADS lesion descriptions and BI-RADS final assessments for each lesion. These lesion descriptors provided by the radiologists were then used as supplemental inputs to our computer technique that was based on computer-extracted image features, to determine the effect of the BI-RADS lesion descriptors on improving the computer performance. We have found that combining the BI-RADS lesion descriptors and computer-extracted image features tend to improve computer classification performance. For

example, For calcifications, the A_z value for the computer performance increased from 0.73 based on computer-extracted image features alone to 0.75, 0.81, and 0.84 based on a combination of the computer-extracted image features and a radiologist-provided BI-RADS descriptors, depending on the particular radiologist. However, we have observed substantial variability in the Bi-RADS descriptors provided by different radiologists. If we train the computer technique on one radiologist's data and then present the computer technique with another radiologist's data, the improvement in the computer performance from the addition of BI-RADS lesion descriptors diminishes. In one experiment, the average A_z value was 0.75 compared to 0.80 if the computer was fed with data from the same radiologist. This finding is important because it provides some insight into an apparent paradox that the BI-RADS, while supposed to help improve mammography interpretation, has been met with mixed results: It may be that BI-RADS is helpful for computer-aided diagnosis but reader variability diminishes this positive effect. Some of this research has been presented at various meetings, including the RSNA [1] and the Army Era of Hope Conference in 2002 [2]. The new results on the effect of reader variability will be presented at the 2004 SPIE Medical Imaging Conference, and we plan to prepare a manuscript for peer-reviewed publication.

Last year we reported on the development of a computer technique for predicting the BI-RADS lesion descriptors that would be selected by a radiologist. The goals are to help streamlining radiologists' reporting on mammogram interpretation by suggesting a detailed draft report, and to help improve the computer classification performance by providing BI-RADS lesion descriptors without requiring a radiologist to do so. We have made further refinement of the method. However, a major challenge for this work is the lack of a "gold standard" for the BI-RADS lesion descriptors that a radiologist would select. BI-RADS lesion descriptors are subjective and descriptive in nature; a "gold standard" for it is neither completely meaningful, nor can be practically assigned. However, its lack thereof presents fundamental problems to the training of our computer classifiers and to the evaluation of same computer classifiers. Given the observation of reader variability in providing BI-RADS lesion descriptors, this challenge is probably difficult to overcome. This work was presented in part at the 2003 SPIE Medical Imaging Conference [3].

2. An analytical comparison of four methods for combining multiple sources of diagnostic information

This particular work is within the scope of SOW Task 2. Computer-aided diagnosis (CAD) methods often analyze each view of a mammogram separately, even for multiple images of the same patient such as the mediolateral oblique and craniocaudal views in mammography [4]. This approach helps simplify the computer technique and generally makes that technique more reliable. However, there is often a need to combine these analyses of multiple images of the same patient to render a result that is clinically relevant. We are investigating several simple methods that have been used in CAD techniques, such as taking the simple average, or taking the result of the one image that is the most indicative of a disease outcome (e.g., malignancy) [4, 5]. We have performed an analytical analysis based on the binormal model for receiver operating characteristic (ROC) analysis with two simplifying assumptions. One assumption is that diagnostic information derived from the multiple images can be described by the same binormal ROC curve; the second assumptions is that the diagnostic information derived from the multiple images is uncorrelated. We have found that the method of simple average always produces an improved ROC curve over the individual image. However, the method of taking the result of the one image that is the most indicative of malignancy and even the method of taking the result of the one image that is the least indicative of malignancy can also improve the ROC curve and, under certain conditions, even outperform the method of simple average to become the preferred method. Based on this analysis, we are able to identify the most appropriate choice of method given the binormal ROC curve parameters. These findings are expected to help improve various CAD methods. This work was presented at the Medical Image Perception Society Conference X and will be presented at the RSNA in 2003 [6, 7]. We will also submit a manuscript of this work for peer-reviewed publication shortly. We are currently generalizing this research by considering correlation in the diagnostic information derived from the multiple images and by considering the situation in which the diagnostic information derived from the multiple images must be described by different ROC curves.

3. Analysis of the influence of radiologist input on the performance of computer classification of malignant and benign calcifications

This particular work is within the scope of SOW Task 3. Previously, in developing a computer technique to classify calcifications in mammograms as malignant or benign, we manually indicated the location of all individual calcifications to the computer and found that the computer can be more accurate than radiologists [4, 8]. In this study, we investigated whether radiologists can be asked to provide minimal input to the computer and obtain consistent computer classification results. Radiologists were instructed to draw a rectangle that encloses all calcifications, and indicate an approximate number of the calcifications (either <6, 6–10, 10–30, or >30). The computer then used these two pieces of information to detect the individual calcifications and, subsequently, classify the calcifications as malignant or benign based on only those calcifications detected by the computer [9]. We showed at the 2002 RSNA conference in an Educational Exhibit 18 cases of digitized mammograms on a computer monitor together with standard and magnification view film mammograms to 38 self-reported breast-imaging radiologists (12 of whom read all 18 cases) [10]. The standard deviation in the location of their rectangles (averaged over all cases) was approximated 3 mm, the standard deviation in the linear dimension of the rectangles was 6 mm, and the standard deviation in the computer-estimated likelihood of malignancy was 17%. These results indicate that it is possible to carry out computer classification on the basis of radiologists' minimal input. This work was presented at the 2003 SPIE Medical Imaging Conference [11] and it formed an important basis for our research on full-field digital mammograms described next.

4. Computer-aided diagnosis of malignant and benign calcifications in full-field digital mammograms

This particular work is within the scope of SOW Task 3. Previously we have developed a computer-aided diagnosis technique to classify breast calcifications as malignant or benign for use on film screen mammograms in which we must identify individual calcifications manually. We have now performed a study to evaluate this technique on full-field digital mammograms. This study represents a totally independent test of the algorithm developed for film screen calcifications on a new, full-field digital mammogram database, with automatic detection of the calcifications by the computer aid. We

analyzed 49 consecutive full-field digital mammograms (29 cancers) showing suspicious calcifications that were biopsied between May 2002 and May 2003 at the University of Chicago. Four mammography specialists read the images retrospectively on a monitor in random order and electronically marked the region of calcifications in each image by making a box around the group. The computer then automatically detected calcifications within the box and analyzed 8 features of calcification morphology and distribution to arrive at an estimate of the likelihood of malignancy. The radiologists entered BI-RADS assessments before and after seeing the computer calculation. Despite variability in input from the radiologists (region selection), the computer achieved consistently high performance with ROC curve areas of 0.80, 0.80, 0.78, and 0.77 (not statistically different). In addition, the average ROC curve area of the unaided observers increased from 0.72 to 0.76 with the computer aid (not statistically significant for the number of cases and observers). Previous testing on film screen mammograms showed the computer aid was able to achieve virtually the same ROC curve area (0.80) as on the digital images and was able to improve radiologists' performance significantly. We conclude from this study that our computer technique can achieve consistently high performance in classifying malignant and benign calcifications in digital mammograms. The consistency and reliability of this technique is important because it was *developed on film* images and *tested on digital* images, without modification. This test simulated clinical use of the technique by including variation in reader input, while demonstrating consistent computer calculation. We have submitted an abstract to the 2004 American Roentgen Ray Society Annual Meeting and we are in the process of preparing a manuscript on this work for peer-reviewed publication.

5. A new method for training artificial neural networks to approximate the ideal observer

This particular work is beyond but related to SOW Task 3. We have continued working on the investigation of some basic properties of artificial neural networks, which we employ in our computer-aided diagnosis methods for malignant and benign breast lesions [8]. While this is not an original Task of the Army program, we are excited about this project and have devoted some effort into it. Insights gained from this project will be used to improve our computer-aided diagnosis technique. We report one significant finding. Artificial neural networks (ANNs) are frequently used in computer-aided diagnosis methods. Generally they are to approximate the ideal observer for some specific classification

task. This is possible given large enough training data. However, unlike the ideal observer, the ANN is not given the full joint probability density function. Instead, it is given the association of a set of training data (input vectors) with one classification outcome, and the association of a second set of training data with a second classification outcome. Often, the ANN is able to approximate the ideal-observer performance; therefore indicating it is able to estimate the full joint probability density function from the information of the training cases. We have developed a method that provides more information to the ANN to help it better approximate the ideal observer. Instead of providing a binary class association for the training data, we provide multiple rank-ordered associations of the training data to classification outcomes. In the extreme of infinite number of such associations, the information provided becomes the posterior probability. For practical implementation, the information provided may be based on, e.g., size or histological grades of tumor when the classification task is to differentiate malignant and benign lesions. The benefits are less statistical variation in the ANN output and better approximation of the ideal observer with limited training data. This work was presented at SPIE Medical Imaging 2003 [12], the 45th Annual Meeting of the AAPM [13], and Medical Image Perception Conference X [14]. We are in the process of preparing a manuscript of this work for peer-reviewed publication.

KEY RESEARCH ACCOMPLISHMENTS

- Determined that the combination of BI-RADS lesion descriptors provided by radiologists and image features extracted by a computer can improve the performance of computer classification of malignant and benign breast lesions in mammograms, but reader variability in providing the BI-RADS lesion descriptors can diminish that improvement.
- Demonstrated that the method of choice for simple un-weighted linear combinations of diagnostic information derived from multiple sources such as multiple images of the same patient is not always a single method but will change from one method to another depending on the ROC curve parameters of the diagnostic information derived from each single source.
- Determined that variability in certain minimal information provided by radiologists is reasonably small in querying our computer-aided diagnosis technique about calcifications in mammograms and

that the resulting variability in computer-calculated likelihood of malignancy for the calcifications is also reasonably small.

- Demonstrated that our computer-aided diagnosis technique for malignant and benign calcifications in mammograms developed on digitized screen-film mammograms that required manual identification of individual calcifications can achieve virtually the same highly accurate and highly consistent performance on full-field digital mammograms, being tested on an independent new database without re-tuning the technique and no-longer requiring manual identification of the individual calcifications.
- Developed a novel technique for training artificial neural networks to better approximate the ideal observer in two-class classification tasks by using multiple training target values instead of the conventional binary training target values.

REPORTABLE OUTCOMES

Manuscripts

1. Jiang Y. Uncertainty in the output of artificial neural networks. IEEE Trans Med Imaging 22:913-921, 2003.
2. Salfity MF, Nishikawa RM, Jiang Y, Papaioannou J. The use of a priori information in the detection of mammographic microcalcifications to improve their classification. Med Phys 30:823-831, 2003.
3. Vyborny CJ, Kukec C, Jiang Y, Doi K. Experience with computer-aided detection in a low-volume mammography clinic. In: Digital Mammography 2002 (Peitgen HO, eds.). Heidelberg: Springer Verlag Publishers, pp. 387-390, 2002.
4. Salfity MF, Nishikawa RM, Jiang Y, Papaioannou J. Improvement in the automatic detection of individual microcalcifications to integrate a cluster-detection and a cluster-classification schemes. In: Digital Mammography 2002 (Peitgen HO, eds.). Heidelberg: Springer Verlag Publishers, pp. 411-413, 2002.
5. Jiang Y, Salfity MF, Chen V, Nishikawa RM, Papaioannou J, Edwards AV, Paquerault S. Effect of radiologists' variability on the performance of computer classification of malignant and benign calcifications in mammograms. Proc SPIE 5034:42-47, 2003.

6. Liu B, Jiang Y. Training artificial neural networks (ANNs) with multiple target values to reduce output uncertainty. Proc SPIE 5034:433-438, 2003.
7. Paquerault S, Jiang Y, Nishikawa RM, Schmidt RA, D'Orsi CJ, Vyborny CJ. Automated selection of BI-RADS lesion descriptors for reporting calcifications in mammograms. Proc SPIE 5032:802-809, 2003.
8. Freedman M, Lo S-CB, Osicka T, Zhao H, Lure F, Xu X, Lin J, Zhang R, Jiang Y. Enhanced computer aided detection of lung cancer on chest radiographs with the Deus Technologies RS-2000. In: CAR'02 Computer Assisted Radiology and Surgery (Lemke HU, Vannier MW, Inamura K, eds.). Amsterdam: Elsevier, 2002.

Abstracts

9. Jiang Y, M. NR, Giger ML, Papaioannou J, Lan L, Vyborny CJ, et al. On-line demonstration of computer-aided diagnosis (CAD) of malignant and benign breast lesions (abstract: educational exhibit). Radiology 225(P):683, 2002. Presented at the 88th Scientific Assembly and Annual Meeting of the Radiological Society of North America (RSNA).
10. Liu B, Jiang Y. Training artificial neural networks (ANNs) with multiple target values for two-class classification problems (abstract). Medical Physics 30:2003. Presented at the 45th Annual Meeting of the American Association of Physicists in Medicine (AAPM).
11. Paquerault S, Jiang Y, Yarusso LM, Papaioannou J, Nishikawa RM. Potential improvement in computerized classification of malignant mammographic clustered microcalcifications using a novel segmentation method (abstract). Medical Physics 30:2003. Presented at the 45th Annual Meeting of the American Association of Physicists in Medicine (AAPM).

Presentations

12. Jiang Y, Liu B. Training artificial neural network with multiple target values to approximate the ideal observer. Presented at Medical Image Perception Conference X, Durham, NC, September, 2003.
13. Liu B, Metz CE, Jiang Y. Proper use of multiple images of the same patient in computer-aided diagnosis (CAD) based on considerations of ROC analysis. Presented at Medical Image Perception Conference X, Durham, NC, September, 2003.

14. Zur R, Jiang Y. Obtaining ideal observer Az value by training ANNs with jitter. Presented at Medical Image Perception Conference X, Durham, NC, September, 2003.

15. Jiang Y, Nishikawa RM, Schmidt RA, Toledano AY, Doi K, Beiden SV, Wagner RF, Campbell G, Metz CE. The Potential of Computer-Aided Diagnosis (CAD) to Reduce Variability in Radiologists' Interpretation of Mammograms. Presented at Biomedical Imaging Research Opportunities Workshop (BIROW), Washington DC, January 2003.

CONCLUSIONS

We have made significant progress toward completing all three SOW Tasks. Tasks 1 and 2 are near completion and we will concentrate on preparing publications. Task 3 is partially completed; we will further improve and test our computer classification technique on full-field digital mammograms and will work on preparing publications. The results support project continuation.

REFERENCES

1. Jiang Y, Schmidt RA, D'Orsi CJ, Vyborny CJ, Nishikawa RM, Paquerault S. Classification of malignant and benign clustered microcalcifications based on computer-extracted lesion features and radiologist-provided BI-RADS description (abstract). Radiology 225(P):497, 2002.

2. Jiang Y, Paquerault S, Nishikawa RM, Giger ML, Schmidt RA, D'Orsi CJ, Vyborny CJ, Metz CE. Computer-aided diagnosis of malignant and benign breast lesions in mammograms. Era of Hope 2002 Department of Defense Breast Cancer Research Program Meeting Orlando, FL: 2002

3. Paquerault S, Jiang Y, Nishikawa RM, Schmidt RA, D'Orsi CJ, Vyborny CJ. Automated selection of BI-RADS lesion descriptors for reporting calcifications in mammograms. Proc SPIE 5032:802-809, 2003.

4. Jiang Y, Nishikawa RM, Wolverton DE, Metz CE, Giger ML, Schmidt RA, Vyborny CJ, Doi K. Malignant and benign clustered microcalcifications: automated feature analysis and classification. Radiology 198:671-678, 1996.

5. Chan HP, Sahiner B, Helvie MA, Petrick N, Roubidoux MA, Wilson TE, Adler DD, Paramagul C, Newman JS, Sanjay-Gopal S. Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study. Radiology 212:817-827, 1999.
6. Liu B, Metz CE, Jiang Y. Proper use of multiple images of the same patient in computer-aided diagnosis (CAD) based on considerations of ROC analysis. Medical Image Perception Conference X Durham, NC: 2003
7. Liu B, Metz CE, Jiang Y. Proper use of multiple images of the same patient in computer-aided diagnosis (CAD) based on considerations of ROC analysis. RSNA Chicago, IL: 2003
8. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis. Acad Radiol 6:22-33, 1999.
9. Salfity MF, Nishikawa RM, Jiang Y, Papaioannou J. The use of a priori information in the detection of mammographic microcalcifications to improve their classification. Med Phys 30:823-831, 2003.
10. Jiang Y, M. NR, Giger ML, Papaioannou J, Lan L, Vyborny CJ, et al. On-line demonstration of computer-aided diagnosis (CAD) of malignant and benign breast lesions (abstract: educational exhibit). Radiology 225(P):683, 2002.
11. Jiang Y, Salfity MF, Chen V, Nishikawa RM, Papaioannou J, Edwards AV, Paquerault S. Effect of radiologists' variability on the performance of computer classification of malignant and benign calcifications in mammograms. Proc SPIE 5034:42-47, 2003.
12. Liu B, Jiang Y. Training artificial neural networks (ANNs) with multiple target values to reduce output uncertainty. Proc SPIE 5034:433-438, 2003.
13. Liu B, Jiang Y. Training artificial neural networks (ANNs) with multiple target values for two-class classification problems (abstract). Medical Physics 30:2003.
14. Jiang Y, Liu B. Training artificial neural network with multiple target values to approximate the ideal observer. Medical Image Perception Conference X Durham, NC: 2003

LIST OF ATTACHED REPRINTS

1. Jiang Y. Uncertainty in the output of artificial neural networks. IEEE Trans Med Imaging 22:913-921, 2003.
2. Salfity MF, Nishikawa RM, Jiang Y, Papaioannou J. The use of a priori information in the detection of mammographic microcalcifications to improve their classification. Med Phys 30:823-831, 2003.
3. Vyborny CJ, Kukec C, Jiang Y, Doi K. Experience with computer-aided detection in a low-volume mammography clinic. In: Digital Mammography 2002 (Peitgen HO, eds.). Heidelberg: Springer Verlag Publishers, pp. 387-390, 2002.
4. Salfity MF, Nishikawa RM, Jiang Y, Papaioannou J. Improvement in the automatic detection of individual microcalcifications to integrate a cluster-detection and a cluster-classification schemes. In: Digital Mammography 2002 (Peitgen HO, eds.). Heidelberg: Springer Verlag Publishers, pp. 411-413, 2002.
5. Jiang Y, Salfity MF, Chen V, Nishikawa RM, Papaioannou J, Edwards AV, Paquerault S. Effect of radiologists' variability on the performance of computer classification of malignant and benign calcifications in mammograms. Proc SPIE 5034:42-47, 2003.
6. Paquerault S, Jiang Y, Nishikawa RM, Schmidt RA, D'Orsi CJ, Vyborny CJ. Automated selection of BI-RADS lesion descriptors for reporting calcifications in mammograms. Proc SPIE 5032:802-809, 2003.

Uncertainty in the Output of Artificial Neural Networks

Yulei Jiang

Abstract—**Analysis of the performance of artificial neural networks (ANNs) is usually based on aggregate results on a population of cases. In this paper, we analyze ANN output corresponding to the individual case. We show variability in the outputs of multiple ANNs that are trained and "optimized" from a common set of training cases. We predict this variability from a theoretical standpoint on the basis that multiple ANNs can be optimized to achieve similar overall performance on a population of cases, but produce different outputs for the same individual case because the ANNs use different weights. We use simulations to show that the average standard deviation in the ANN output can be two orders of magnitude higher than the standard deviation in the ANN overall performance measured by the A_z value. We further show this variability using an example in mammography where the ANNs are used to classify clustered microcalcifications as malignant or benign based on image features extracted from mammograms. This variability in the ANN output is generally not recognized because a trained individual ANN becomes a deterministic model. Recognition of this variability and the deterministic view of the ANN present a fundamental contradiction. The implication of this variability to the classification task warrants additional study.**

Index Terms—**Artificial neural networks, classification, computer-aided diagnosis, estimation uncertainty, prediction error.**

I. Introduction

ARTIFICIAL neural networks (ANNs) are frequently used to perform classification tasks in medical imaging applications, e.g., computer-aided diagnosis (CAD) [1], [2]. In general, ANNs represent families of mathematical formulas that combine and transform an input data vector into a quantitative output or outputs. The parameters of these formulas are determined in an iterative training process in which the parameters are adjusted in an attempt to match the output produced from a set of training cases to target output values. ANNs usually take the form of multiple nodes in successive layers. The adjustable parameters are represented by weights that connect the nodes in adjacent layers, modulated by some activation function that can be nonlinear. One common application of ANNs is as pattern classifiers because ANNs can be trained to recognize patterns in a set of training cases and then match unknown cases to the patterns to perform a classification task. This is attractive because it is accomplished without someone explicitly designing a mathematical form for the classifier. In CAD, ANNs are used to classify true lesions from computer-identified false positives [3], [4], malignant from benign lesions [5]–[8], or one from several other differential diagnoses [9]. The typical way in which the ANN is employed is that, first, other methods are used to extract from the image features of the object of interest, such as morphological properties, and second, the ANN is used to generate a classification decision or prediction. Being the classifier at the final stage of a computer technique, the ANN plays a key role in determining the overall performance of the computer technique. In addition, some CAD techniques require humans (radiologists) to interpret the ANN output, thereby giving the ANN a whole new role to affect system performance [7]–[9]. Optimization of the ANNs and a thorough understanding of the statistical properties of the ANN output are, therefore, imperative for CAD applications.

ANN optimization is limited in practice by a finite training case sample and is accomplished through a stochastic training process. This stochastic process gives ANN the ability to avoid being trapped at local minima. At the same time, this stochastic process makes ANN optimization empirical and subject to strong influence from statistical variations. It has been shown that performance of the properly optimized ANN and one's ability to measure that performance accurately depend on the numbers of training and testing cases, and how one uses a fixed set of cases in training and test (i.e., resampling plans) [10]–[12]. These works focus on the overall performance of the ANN on a population of cases. In this paper, we report on the statistical variation in the ANN output on the individual case. We demonstrate statistical variability in the ANN output and show that this variability is larger than the variability in the ANN overall performance.

II. Theory

A. Variability in the ANN Output

Conventionally, one develops a single ANN through training and, once completed, holds the ANN parameters fixed. Such an ANN becomes a deterministic model that a given input data vector will always produce a particular, completely predictable output value. Therefore, if this ANN is applied repeatedly to a given image, the ANN output will always be the same (within numerical precision of the computer). This phenomenon gives rise to an illusion that the ANN is equivalent to a deterministic mathematical formula in that its outputs are infinitely reproducible (or at least up to the limit of the numerical precision of the computer used to implement the ANN).

Manuscript received September 13, 2002; revised January 22, 2003. This work was supported in part by National Cancer Institute (NCI)/National Institutes of Health (NIH) under Grant R21 CA93989 and in part by the U.S. Army Medical Research and Materiel Command under Grant DAMD17-00-1-0197. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was J. Liang.

The author is with the Kurt Rossmann Laboratories for Radiologic Image Research, Department of Radiology, MC2026, The University of Chicago, 5841 South Maryland Ave., Chicago, IL 60637 USA (e-mail: y-jiang@uchicago.edu).

Digital Object Identifier 10.1109/TMI.2003.815061

This deterministic depiction of the ANN is flawed because it ignores, in most practical situations, the stochastic nature of ANN training—the process that determines the seemingly deterministic mathematical formula the ANN uses to calculate its output. In the limit of an infinitely large training dataset, training of the ANN should normally approach an infinitely large number of epochs that, if we assume trapping at local minima does not occur, guarantees the result of an optimal ANN. However, with a finite training dataset, it is usually not desirable to train the ANN indefinitely because of possible over-fitting [13]. Unlike in the case of an infinitely large training dataset, an asymptotic solution (i.e., a set of weights) for the ANN does not exist in such situations, and it is usually best to stop the training process at some point that is determined empirically—and sometimes rather arbitrarily as we describe next.

The choice of the one ANN from all others or, equivalently, the choice of the number of training epochs, can be made based on the overall performance of the ANN that can be characterized by the receiver operating characteristic (ROC) curve [14], [15]. Therefore, there is a clear reason to choose one particular ANN with a higher ROC curve over another ANN with a lower ROC curve. However, this choice becomes rather arbitrary when the ROC curves of the ANNs are similar. Typically, as the number of training epochs increase, the overall performance of the ANN follows an increasing but noisy trajectory. As the ANN approaches its "optimal" performance, the overall performance of the ANN becomes relatively constant as the number of training epochs continues to increase. The "final" ANN is usually chosen from these ANNs with similar overall performance. The performance of the ANN may decrease if the training epochs increase further, when over-fitting occurs.

Given the lack of an asymptotic ANN in most practical situations, a question arises concerning the variability in the ANN that one chooses to use. It is clear that once one chooses an ANN to use, the output from that ANN bears no variability because it will always produce the same predictable output for a given ANN input data vector. However, in the absence of an asymptotic ANN, one could very well have chosen a different ANN with a different set of weights. This second ANN will also bear no variability because it, too, will always produce the same predictable output for a given ANN input data vector. However, the outputs from the first and the second ANN with respect to the same input data vector need not be the same. In general, one would expect these ANN outputs to be different because the ANNs have different sets of weights, i.e., they use different mathematical formulas to calculate their outputs. Therefore, the issue of ANN output variability arises given that there is no particular reason to choose one ANN over another.

This discussion of the variability in the ANN output will not be meaningful for arbitrarily selected ANNs that have disparate overall performance because for all practical purposes low performers are not of interest. However, for ANNs that have similar overall performance that makes it arbitrary to choose one over another, variability in the ANN outputs should not be ignored. Therefore, one can additionally ask whether the variability in ANN outputs is greater than the variability in the ANN overall performance for those ANNs that can be legitimately considered as "optimized."

B. *Training of Multiple ANNs*

To investigate the variability in ANN outputs, one needs to obtain multiple "optimized" ANNs based on a given training dataset. In this paper, we obtained these ANNs by assigning different "seed values" to a pseudo random number generator that regulates the ANN training process. This seed value determines the sequence of random numbers that were used to determine the initial weights of the ANN and the training case sequence within each individual training epoch. Therefore, given the training dataset, this seed value to the random number generator determines the training trajectory in the ANN weight space. By varying this seed value, one in general forces the ANN to follow a different training trajectory.

To determine the "optimal training epochs," we used an independent test dataset to monitor the overall performance of the ANNs as they were being trained. We then empirically selected a particular number of training epochs for each ANN based on the ANN performance on this test dataset. Fig. 1(a) shows an example of the performance of eight ANNs measured by the area under the ROC curve, A_z [16], obtained from the test dataset and plotted as a function of the number of training epochs. This graph shows that the performance of the ANNs improved rapidly during the initial training epochs. The ANN performance then became relatively constant as the number of training epochs increased to about 500, before a varying degree of over-fitting occurred as the number of training epochs increased further. The training trajectories of the ANNs are noisy and they are clearly different from each other at least for some portion of the trajectories, e.g., at greater than 500 training epochs. This confirms that our use of different seed values to the random number generator did lead the ANNs to follow different training trajectories.

Based on Fig. 1(a), we selected the "optimal training epochs" to be 70 because at this number of training epochs, all of the ANNs had a very similar overall performance that approaches their best performance. Two ANNs showed slightly higher A_z values at a greater number of training epochs; but not all the ANNs achieved the slightly better performance. Therefore, the occasional slightly better performance was considered a random event that may be a consequence of the finite size of the test dataset or the uncertainty in the fitting of the ROC curves and do not represent the "optimal" performance of the ANNs. Although in general the ANNs are not expected to converge to the "optimal" performance at a common number of training epochs as they generally follow different training trajectories, Fig. 1(a) shows that these ANNs converged to the "optimal" performance at a fairly homogeneous speed. Therefore, for simplicity, we chose a single number of "optimal training epochs" for all ANNs in this paper.

The A_z value is a widely used summary index for ROC curves. However, it is possible for ROC curves of different shapes that represent different ANN performance to have the same A_z value [17]. To ensure that the "optimal" ANN performance we chose based on the A_z value indeed represented a common performance or a common ROC curve shared by all ANNs, we also plotted the binormal ROC curve parameters, $\hat{a}$ and $\hat{b}$ (the carets indicate that these are estimated parameters),

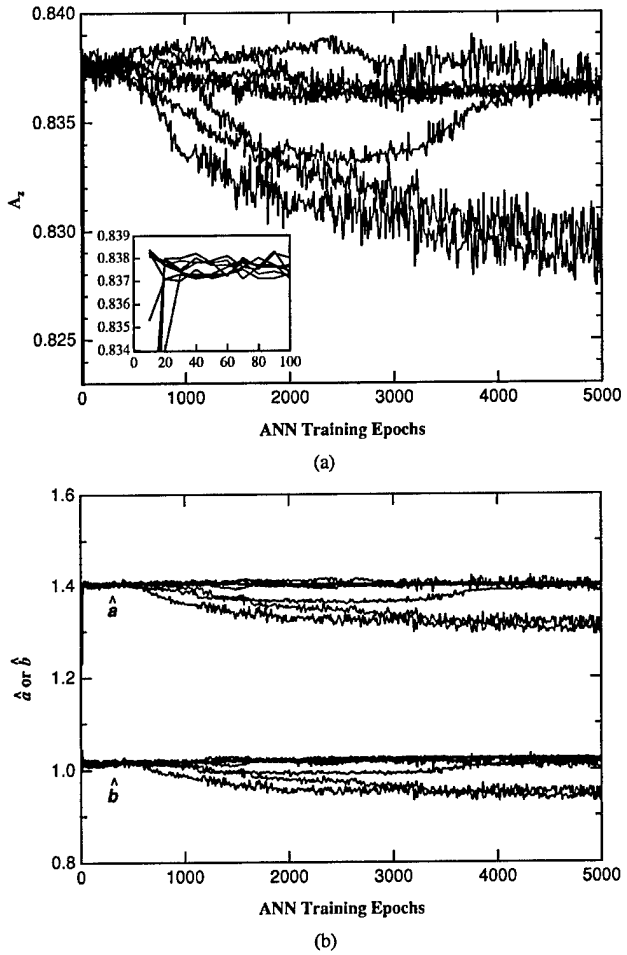


Fig. 1. Variation of (a) the training trajectory measured by A_z and (b) the $\hat{a}$ and $\hat{b}$ parameters of fitted binormal ROC curves (the carets indicate that these are estimated parameters) of eight ANNs trained on the basis of a single training dataset. These trajectories show that the multiple ANNs that we trained achieved highly similar overall performance at about 70 epochs. The ANNs were trained with arbitrary but different seed values to a random number generator and otherwise with identical parameters. Insert shows details of the first 100 epochs. The ANNs had a 2-2-1 structure. The simulated datasets were generated with parameters of $a_o = 1$ and $b_o = 1$. Note that $\hat{a}$ is different from a_o because $\hat{a}$ is an estimate of the Euclidian distance between the means of the estimated class distributions in 2 dimensions whereas a_o is a one-dimensional parameter.

as a function of the number of training epochs. An example is shown in Fig. 1(b). It is apparent that between 10 and 500 training epochs, the ROC curve parameters are virtually the same for all the ANNs and, hence, their very similar A_z values.

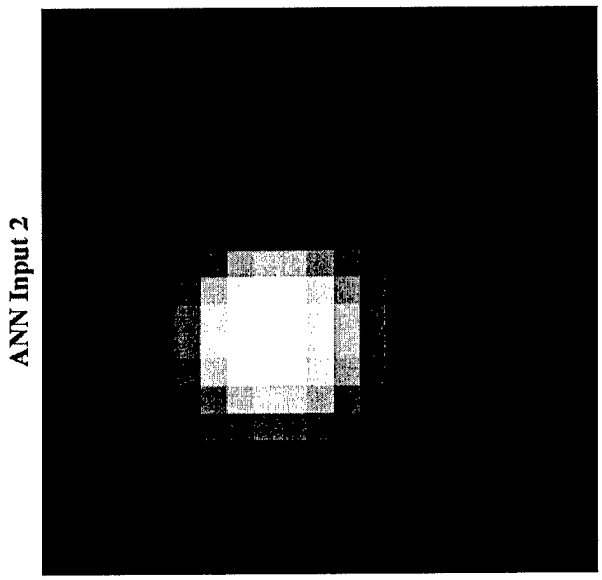
III. Methods

A. Artificial Neural Networks

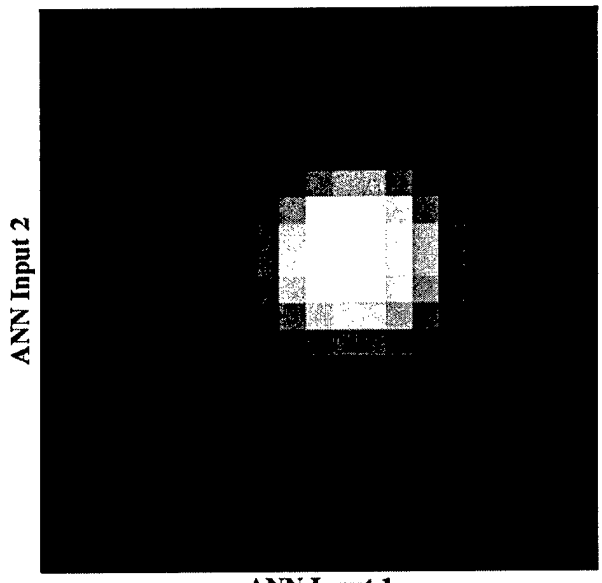
We used feedforward-error back propagation ANNs in this paper [18]. The typical ANN had three layers: the input layer had the same number of nodes as the dimension of the ANN input data vector; the hidden layer had a variable number of nodes that was set empirically; and the output layer had one node. The nodes in adjacent layers were fully connected. Although in principle the multiple ANNs trained on the basis of a given training dataset do not need to have the same network structure, i.e., the same number of nodes in each layer, for simplicity we only used ANNs of the same network structure in this paper. Because these ANNs did achieve very similar and near optimal overall performance, as evident from Fig. 1, we believe that this simplification did not affect the results of this paper. All other empirically determined parameters including a learning rate and a weight bias were also kept the same for the ANNs except for the random number generator seed value that was different for each ANN. All values of the ANN input data vector were normalized to between zero and one. The training and the test input data shared the same normalization factors, and each dimension of the input data vector was normalized independently. The binary values of 0.1 and 0.9 were used as the ANN target output values in training the ANNs to help improve convergence in ANN training.

B. Simulation Study

We used both simulated datasets and a dataset from a mammography application in this paper. In the simulation study, we assumed that the two distinct classes the ANNs were designed to distinguish follow multivariate and isotropic normal distributions. One class may represent, e.g., the normal or disease-free cases, or in a different classification problem, benign lesions, whereas the other class represents the abnormal or diseased cases, or malignant lesions. The dimensionality of the distributions, i.e., the number of variables in the ANN input data vector was chosen to be 2 and 8, hereafter referred to as two-dimensional (2-D) datasets and eight-dimensional (8-D) datasets. We started with 2-D datasets because it is the simplest classification problem for an ANN and because one can visualize 2-D data relatively easily. We used 8-D datasets because the ANN input data vector in our mammography dataset was also 8-D. Under the normality and isotropy assumptions, following a common convention used in ROC analysis and without loss of generality, we assumed that the normal class follows the multivariate standard normal distribution and that the abnormal class follows an isotropic multivariate normal distribution with mean of $\mathbf{u}$, where $u_i = a_o/b_o$ for all i, and an isotropic variance of $1/b_o$. Therefore, for a 2-D dataset, the distribution that represents the normal class centers at (0, 0) with an isotropic variance of one, and the distribution that represents the abnormal class centers at $(a_o/b_o, a_o/b_o)$ with an isotropic variance of $1/b_o$. For $a_o = 1$ and $b_o = 1$, the abnormal distribution centers at (1, 1) with an isotropic variance of one. The assumption that the mean of the abnormal distribution falls on the 45° line does not cause a loss of generality because all other locations of the mean can be transformed to the 45° line through axis rotation. Under the normality assumption, the ideal observer who uses the likelihood ratio defined by the class distributions as the decision variable follows a linear decision boundary (or a hyper plane in higher dimensions) that is perpendicular to the 45° line. With $b_o = 1$, the ideal observer's ROC curve is obtained by sweeping this linear decision boundary monotonically from $-\infty$ to $+\infty$. For $b_o \neq 1$, parts of the ideal observer's ROC curve correspond to two parallel linear decision boundaries. For easy visualization of the data, we used only $b_o = 1$ in this simulation study. For



(a)



(b)

Fig. 2. The presumed distributions of (a) the normal class and (b) the abnormal class in a 2-D simulation study for $a_o = 1$ and $b_o = 1$.

$b_o = 1$, the ideal observer's A_z value is determined only by the Euclidian distance between the means of the class distributions. For $a_o = 1$, the ideal observer's A_z value is 0.84. We used the following parameters in the simulation study chosen arbitrarily and in part because the corresponding ideal observer's A_z covered a wide range of values: for 2-D datasets, $a_o = 0.707$, 1, 1.202, and 1.414 with the ideal observer's A_z values of 0.76, 0.84, 0.89, and 0.92, respectively; and for 8-D datasets, $a_o = 0.707$, 1, and 1.202 with the ideal observer's A_z values of 0.92, 0.98, and 0.99, respectively. An example of the 2-D distributions is shown in Fig. 2 for $a_o = 1$ and $b_o = 1$.

For ANN training, we used training datasets of 200 cases or 1000 cases with an equal number of cases from each class drawn randomly from their respective distributions. For ANN testing, we used an independent test dataset of 2000 cases with an equal number of cases from each class drawn randomly from the respective distributions. For the 2-D datasets, the ANN had two hidden nodes and, therefore, a 2-2-1 network structure. For the 8-D datasets, an 8-6-1 network structure was used. Eight ANNs were trained on the basis of each training dataset. Each ANN was trained using a different seed value for the random number generator, set arbitrarily to be 1001, 2001, 3001, etc. Training of the eight ANNs was stopped at a common number of training epochs based on test data similar to those shown in Fig. 1.

C. Mammography Study

A similar study using a mammography dataset was carried out. The purpose was to demonstrate an example of the similar effects as observed in the simulation studies. The task of this mammography application was to classify clustered microcalcifications in mammograms as malignant or benign based on eight computer-extracted image features. This application is described in detail elsewhere [19]. An 8-6-1 network structure was used for the ANNs (the number of hidden nodes, 6, was determined as appropriate in a previous work [19]). The dataset contained 53 cases (19 malignant) and a total of 107 individual mammograms (40 depicting malignant microcalcifications). A leave-one-out method was used because of the small size of this dataset [10]. Using the leave-one-out method, the dataset was partitioned such that 52 cases were used as training cases and the one left-out case was used as a test case. Once the training and test process was completed, the dataset was re-partitioned with the same number of training and test cases but with a different test case. This process was repeated until all 53 cases were used as a test case and, subsequently, an ROC curve and an A_z value were computed from the results of all 53 test cases. The advantage of this method is that all 53 cases can be used for both training and testing while the test cases are independent from the training dataset to avoid a learning bias. Because a typical case consisted of two mammograms of the same cluster of microcalcifications imaged from the mediolateral oblique and the crandiocaudal projections, these mammograms from the same patient were kept as a unit when the case was assigned as either a training case or a test case, so that the test cases were truly independent of the training cases. However, the ANNs analyzed each mammogram independently even though it may be one of two films from the same case. These ANN outputs from the mammograms of the same case were combined into a single number by retaining only the maximum output value (i.e., the output most indicative of a malignancy) for calculation of the A_z values [19]. Eight ANNs were trained with identical parameters except for different seed values to the random number generator that were arbitrarily set to be 1001, 2001, 3001, etc.

IV. Results

A. Simulation Study of 2-D Datasets

A comparison of the variability in the overall ANN performance and the variability in the individual ANN output is shown

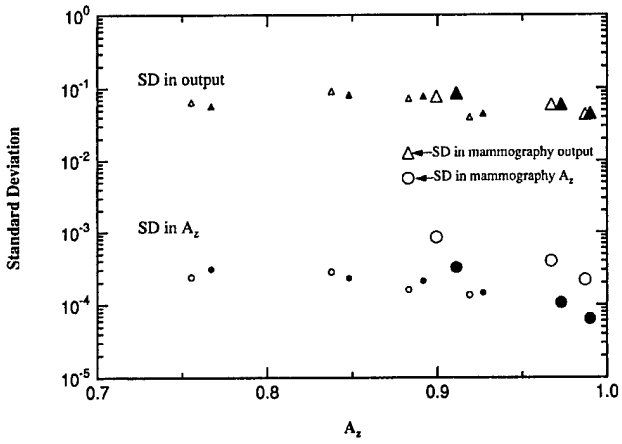


Fig. 3. Comparison of the magnitude of the standard deviation in A_z and the magnitude of the standard deviation in ANN output. The standard deviations in A_z are considerably larger than the standard deviations in ANN output. Standard deviation in ANN output was calculated for each given test case and was averaged over 2000 test cases. Data were obtained from eight ANNs trained on the basis of a given training dataset, with identical parameters except for an arbitrary but different seed value to a random number generator. ANN structures were 2-2-1 (small symbols) or 8-6-1 (large symbols). The number of training cases was 200 (hollow symbols) or 1000 (solid symbols) with an equal number of training cases from each class. For mammography results, the total number of cases was 53 and a leave-one-case-out method was used for training and for performance evaluation.

in Fig. 3. Each data point was based on eight ANNs developed from a given training dataset. On the one hand, the variability in the overall ANN performance was quantified by the standard deviation of the eight A_z values. Each A_z value represented the performance of one ANN measured from 2000 independent and randomly sampled test cases. On the other hand, the variability in the individual ANN output was represented by the standard deviation of the output values for a given test case. Because there were 2000 test cases, Fig. 3 plots the average of the 2000 standard deviations that correspond to each test case. Each data point corresponds to the results of a common number of training epochs shared by eight ANNs trained on a given training dataset. To minimize the chance of unreliable results, we recorded the results of a total of three numbers of training epochs for each training dataset. These results show virtually identical average A_z values and only small variations in the standard deviations in A_z and in the ANN outputs. Therefore, for clarity, Fig. 3 shows only one data point for each training dataset.

The standard deviation in the A_z value was negligible for all practical purposes (< 0.001). This is consistent with Fig. 1 that shows we were able to obtain multiple ANNs that have very similar, nearly optimized, overall performance. There is a slight downward trend of the standard deviations in A_z as the average A_z value increases. This indicates that as the A_z value increases, it is easier to optimize the ANNs and the performance of the multiple ANNs becomes more similar.

The average standard deviation in the ANN output was on the order of 0.01 to 0.1. Given that the ANN output values are between zero and one, these values are small but are not negligible. The average standard deviation in the ANN output is approximately two orders of magnitude larger than the standard deviation in the A_z values. Therefore, while the overall performance of the ANNs converges toward the optimal performance, the variability in the ANN output for individual test case remains relatively large.

Furthermore, the variability in the ANN output is not a constant for all test cases. Fig. 4(a) shows the standard deviation in the ANN output as a function of the ANN input data vector. Fig. 4(a) plots the input data as a 2-D matrix. Each dimension of this matrix represents one scalar feature value that is scaled to between zero and one for input to the ANN. The two classes that the ANNs were trained to classify populated this matrix as shown in Fig. 2. Fig. 4(a) shows that the variability in the ANN output is smallest at the lower-left and upper-right corners. According to Fig. 2, these two corners are most likely populated by one class only. Therefore, the variability in the ANN output was smallest where only one class occupies the local ANN input data space. The variability in the ANN output increases toward the negative diagonal and it is greatest near the negative diagonal, where the two classes have an equal probability to populate (see Fig. 2).

Fig. 4(a) shows that the variability in the ANN output follows a linear pattern that sweeps across the matrix of the input data approximately in parallel to the negative diagonal. This linear pattern corresponds to the linear decision boundaries manufactured by the ANNs during their training process. Fig. 4(b) shows the average output of the eight ANNs as a function of the ANN input data vector, plotted in the same way as Fig. 4(a). The average ANN output follows the same linear pattern as the variability in the ANN output that sweeps across the matrix of the ANN input data approximately in parallel to the negative diagonal. However, unlike the variability in the ANN output that maximizes near the negative diagonal, the average ANN output increases monotonically from the corner most likely occupied by one class, to the opposite corner most likely occupied by the other class. These decision boundaries are virtually the same as those used by the ideal observer in this simple classification problem, offering yet another confirmation that the ANNs were nearly optimized during their respective training processes.

Fig. 4 shows that the variability in the ANN output was smallest when the average ANN output unequivocally predicts the outcome of one class versus the other, i.e., when the average ANN output is near zero or one. On the other hand, the variability in the ANN output was greatest when the average ANN output was equivocal, i.e., when the average ANN output was about 0.5. This is more clearly shown in Fig. 5 that plots the standard deviation in the ANN output as a function of the average ANN output. In this particular example, the variability in the ANN output follows a tight band that maximizes when the average ANN output is between 0.5 and 0.6.

This dependence of the ANN output variability on the magnitude of the output is likely caused in part by a dependence of the sum-of-square error used in ANN training on the training case ratio. While the total number of training cases from each class is fixed in a given training dataset, the local training caseratio can vary considerably in small regions of the ANN input data space. For example, in the distributions shown in Fig. 2, the local training case ratio is close to unity near the negative diagonal, but is far from unity near the lower-left and upper-right corners. Fig. 6 plots the contribution from each training case to the sum-of-square error as a function of ANN output and as a function of the training case ratio. With equal numbers of training

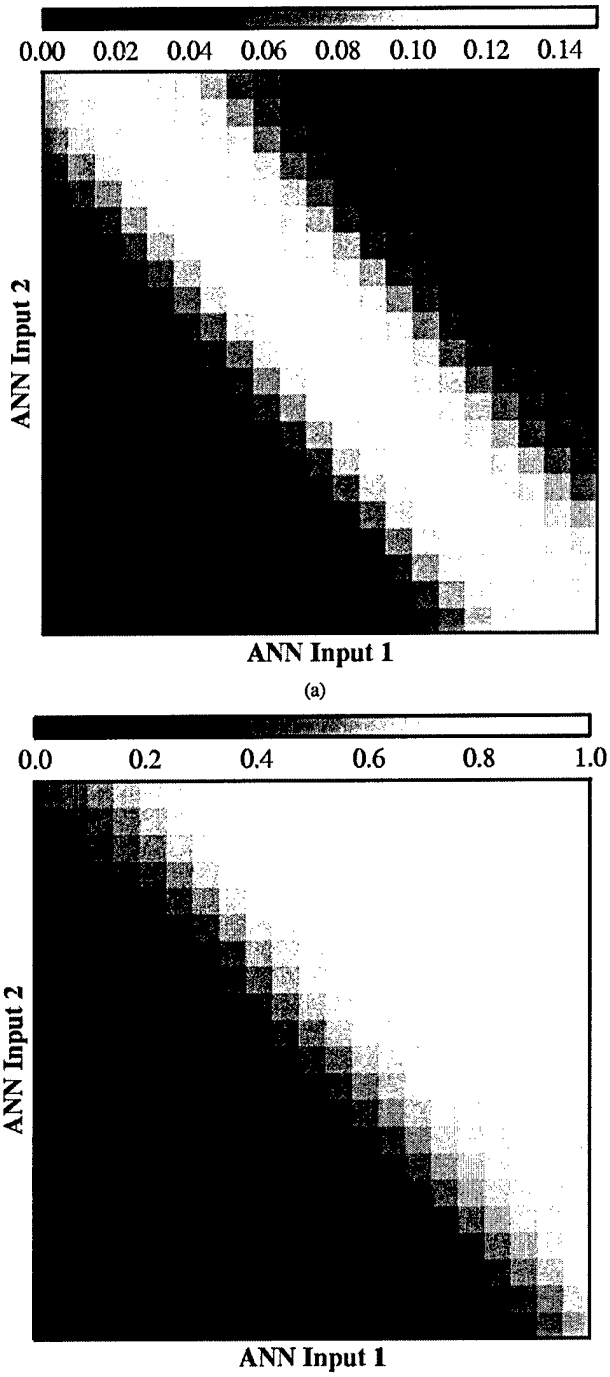


Fig. 4. The distribution of the magnitude of (a) the standard deviation and (b) the average value in the output of eight ANNs plotted in the ANN input data vector space sampled from the distributions shown in Fig. 2 (training cases $n = 200$ with equal number of cases from each class, ANN structure 2-2-1). The standard deviation in the ANN outputs are not uniform in the ANN input data vector space and the average of the ANN outputs shows that the ANNs are similar to an optimal linear classifier. Top bars show grayscale maps. The ANNs were trained on the basis of a single training dataset with identical parameters except for an arbitrary but different seed value to a random number generator.

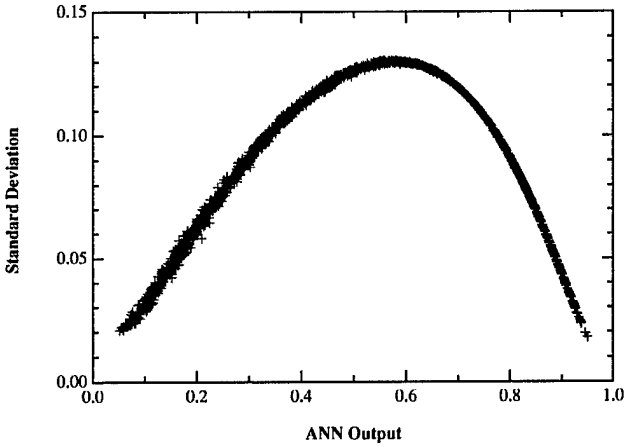


Fig. 5. The standard deviation in the output of eight ANNs as a function of the average ANN output. See also Fig. 4. The ANNs were trained on the basis of a single training dataset with identical parameters except for an arbitrary but different seed value to a random number generator. Training cases ($n = 200$ with equal number of cases from each class) were sampled from the distributions shown in Fig. 2. ANN structure was 2-2-1.

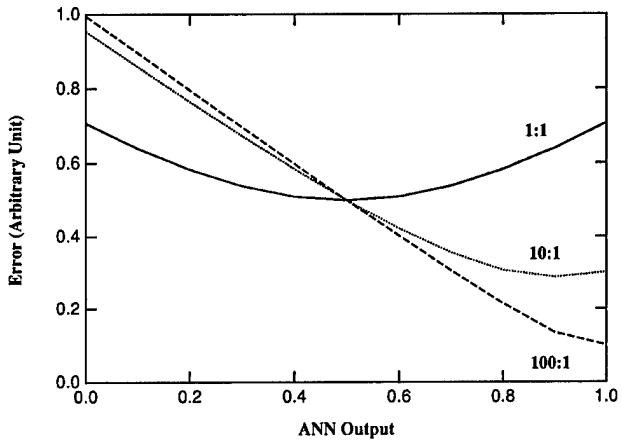


Fig. 6. Schematic illustration of the relationship between the sum-of-square error used in ANN training and the ratio of the training cases from each input data class. Although the overall number of training cases from each class may be equal, this relationship will prevail locally in the vector space of the input data where the local ratio of the training cases from each class can be far from unity.

cases from each class, i.e., for a training case ratio of 1:1, the ANN output value that minimizes the sum-of-square error is 0.5. However, the error curve for a 1:1 training case ratio is relatively shallow. Therefore, an ANN output of 0.8 will produce a relatively small sum-of-square error (0.58 versus 0.50). In contrast, the error curve for a 100:1 training case ratio follows a steeper curve. The ANN output value that minimizes the sum-of-square error is 1.0 and an ANN output of 0.7 will produce a relatively large sum-of-square error (0.31 versus 0.11). The effect of this on the training of multiple ANNs is that in the case of a 100:1 training case ratio, the ANNs will be more likely to produce very similar output values close to one (or zero), whereas in the case of a 1:1 training case ratio, the ANNs will produce output values close to 0.5 but will be less likely to produce the same output value. Note that the fact that the minimum error of 0.5 for a 1:1 training case ratio is considerably greater than the minimum error of 0.11 for a 100:1 training case

ratio. This will exacerbate the variability in the case of a 1:1 training case ratio because the ANN will continue modify its weights to try to reduce the sum-of-square error even when the minimum error of 0.5 has been achieved for a 1:1 training case ratio. Therefore, one would predict based on Figs. 2 and 6 that the multiple ANNs will have small variability in their output values near the lower-left and upper-right corners and will have greatest variability near the negative diagonal. The effect of this relationship between the sum-of-square error and the training case ratio can be greater than Fig. 6 indicates in certain regions of the ANN input data space because of the nonuniform probability of the local total number of cases. Fig. 6 shows this effect normalized to each case and does not show the effect of the (local) total number of cases.

To prove that the local training case ratio contributes to the causes for the dependence of the ANN output variability on the magnitude of the output, we modified the ANN training target values so that instead of using binary training target values, we used the likelihood ratio calculated from the two underlying class distributions as the training target value. By doing so, we eliminated the dependence of the sum-of-square error used in ANN training on the local training case ratio. With this modification, the average standard deviation in the ANN output was reduced by one order of magnitude or more while the standard deviation in the A_z value remained virtually unchanged. The variability in the ANN output no longer had the characteristic dependence on the magnitude of the output of larger variability near the output of 0.5, indicating that the effects of local training case ratio was eliminated. However, with a reduced order of magnitude, the variability in the ANN output was still not a constant for all output values, suggesting that there may be other secondary causes for the dependence of the variability in the ANN output on the magnitude of the output.

B. Simulation Study of 8-D Datasets

Results from the 8-D simulated datasets were similar to those from the 2-D simulated datasets. A comparison of the variability in the overall ANN performance and the variability in the ANN output is shown in Fig. 3 alongside results from the 2-D datasets. As expected, for the same a_o and b_o parameters, the average A_z values from the 8-D datasets were substantially higher than from the 2-D datasets. But the standard deviation in the A_z values of the eight ANNs and the average standard deviation in the ANN outputs were comparable to those of the 2-D datasets. The average standard deviations in the ANN outputs were approximately two orders of magnitude higher than the standard deviations in the A_z values.

C. Mammography Study

Results of the variability in the A_z values and in the ANN outputs from the mammography study are also shown in Fig. 3. The standard deviation in the A_z value was 0.005, larger than those obtained in the simulation studies. The average standard deviation in the ANN outputs was 0.013, smaller than, but within the same order of magnitude of, those from the simulation studies. The average standard deviation in the ANN outputs was approximately twice as large as the standard deviation in the A_z values. These results were obtained from a smaller dataset (53 cases) using a leave-one-out method.

V. Discussion

The purpose of this paper is to show the existence of variability in the ANN output. Using simulations and an example from a mammography application, we found that the average standard deviation in the ANN output is on the order of 0.01 to 0.1. It is larger than the standard deviation in the A_z values that measures the overall ANN performance, and it is two orders of magnitude larger than the standard deviation in the A_z values in the simulations. The magnitude of the standard deviation in the ANN output appears to be large enough to have some practical implications on the use of the ANN outputs and, therefore, needs to be studied, but it is small enough to not undermine seriously the general reliability of ANNs. These findings contrast a common misconception that views the ANN as a deterministic mathematical model without variability. This variability in the ANN output should not be a surprise because the task of statistical prediction or classification is inherently uncertain. In the statistical estimation of a "population value," an estimate from a measurement made on a given sample may be perfectly reproducible, but in general it will not be perfectly reproducible if measured from a different sample. Because of this, all statistical estimates are considered inherently uncertain. Similarly, in the use of ANNs, we are interested in the "best" prediction the ANN can make, not in the prediction of a particular ANN chosen somewhat arbitrarily. Therefore, it is only natural to recognize that the ANN output is associated with statistical variability. What may be surprising is that little or no attention has been devoted to this type of ANN variability in the medical imaging literature [20], [21].

While we compared the magnitude of the ANN output variability to the standard deviation in the A_z value, this comparison is not necessary to recognize the variability in the ANN output, and we used the A_z value simply as a point of reference because it is on a similar numerical scale as the ANN output. One can also compare the magnitude of the ANN output variability to the scale of the ANN output and conclude that the output variability is not negligible. However, it was necessary for us to use the A_z value as a means to ensure that the multiple ANNs that we obtained were close to being optimized because otherwise the analysis would not have been meaningful.

We will explore the practical implications of the variability in the ANN output in a future study. It is clear that the variability in the ANN output does not necessarily affect the overall ANN performance. In another word, the variability in the ANN output that we demonstrate is invisible to ROC analysis. This is because systematic shifts in the ANN output that affect the two classes in the same way will not affect the A_z value, but will be seen as variability in the ANN output. Indeed, multiple ANNs can be trained from a single training dataset to achieve highly similar overall performance as measured by the A_z value while exhibiting relatively large variability in the ANN output. However, the variability in the ANN output may affect the use of the ANN output as a classifier prediction because this variability could cause the ANN output to be interpreted inaccurately, particularly when humans (radiologists) must interpret the ANN output [7], [8].

The nonuniform dependence of the variability in the ANN output on the magnitude of the output as shown in Fig. 5 may further affect the interpretation of the ANN output. Variability in the ANN output tends to be largest where the ANN output is equivocal, further weakening the ANN prediction in such outputs. We have identified one cause for the variability in the ANN output to be the use of binary training target output values and the consequent dependence of the sum-of-square error on the local training case ratio in the ANN input data space. We showed that using the likelihood ratio as the training target output value could reduce the variability in the ANN output because it eliminates the dependence of the sum-of-square error on the local training case ratio. Unfortunately, however, this method cannot be used in real-world classification tasks where the likelihood ratio is unknown.

The findings that we report here are based in large part on simulation studies of simple classification tasks in which the two classes are assumed to follow isotropic multivariate normal distributions. These simple classification tasks were chosen for illustration purposes because a theoretical expectation can be derived *a priori* and because the results are readily interpretable. However, these simple classification tasks may not adequately reflect the behavior of the ANNs in real-world classification tasks. Furthermore, we studied only the type of feed-forward and error back propagation ANN that may not reflect the behavior of other types of ANNs. Moreover, we have adopted a number of simplifications such as fixing the ANN structure and using a common number of training epochs for multiple ANNs trained on the basis of a single training dataset. These may have affected our findings in some unknown way. Nevertheless, despite these simplifications, our semi-theoretical analysis indicates that variability in the ANN output is a natural expectation. We expect that similar variability in the ANN output will be demonstrated in more sophisticated and more realistic simulations, and in the analyses of real-world ANN classifiers.

The results from our mammography study could help put into perspective the findings from our simulations. Like in the simulations, the average standard deviation in the ANN output was larger than the standard deviation in the A_z values in the mammography study. However, unlike in the simulations, the difference between the standard deviations in the ANN output and in the A_z values is much smaller in the mammography study. A careful inspection of Fig. 3 shows that the standard deviation in the A_z values is one order of magnitude larger than those in the simulations, whereas the standard deviation in the ANN output is within the same order of magnitude as those in the simulations. Two reasons may have contributed to this. The first is that the leave-one-out method was used in the mammography study but not in the simulations. Because of the leave-one-out method, an A_z value actually characterizes 53 separate training processes using 53 slightly different training datasets. One would expect the 53 different training datasets to induce greater variability than any single training dataset. In addition, while the substantial similarities among the 53 different training datasets would lead one to expect similar ANN performance, it is not as readily expected that the ANNs would achieve similar performance at the same number of training epochs. However, the leave-one-out method does not account for these differences and uses a common number of training epochs for the 53 ANNs trained on all training datasets. As a result, the magnitude of noise in the training trajectory (i.e., fluctuation) was greater in the A_z values as a function of the number of training epochs. The effect is that it is difficult to obtain multiple ANNs (that were trained with different random seed values) that have very similar A_z values, particularly when we required that the multiple ANNs share a common number of training epochs as in the simulations. The second reason is that the number of cases in the mammography study (53 cases) was smaller than in the simulations. The small number of cases made it necessary to use the leave-one-out method, but the leave-one-out method likely had compounded the effect of the small number of cases to induce greater variability in the A_z values. The standard deviation in ANN output exhibited the characteristic dependence on the magnitude of the output that is larger near the output of 0.5 (as shown in Fig. 5), but this relationship did not conform to a tight band as in the simulations, possibly due to additional sources of variation such as the use of the leave-one-out method and the small number of cases.

In summary, we have shown that the outputs of multiple ANNs trained on the basis of a single training dataset that achieve highly similar overall performance as measured by the A_z value exhibit small but nonnegligible variability. The average standard deviation in the outputs of the multiple ANNs can be two orders of magnitude larger than the standard deviation in the A_z values. The variability in the ANN output is caused in part by the use of binary training target values and a consequent dependence of the sum-of-square ANN training error on the local training case ratio in the ANN input data space. The magnitude and dependence of this variability in the ANN output, and the implication of this variability on the use of the ANN predictive output need to be studied further.

ACKNOWLEDGMENT

The author would like to thank C. E. Metz, Ph.D. for many helpful discussions and B. Liu, Ph.D. for critically reading the manuscript.

REFERENCES

[1] J. M. Boone, G. W. Gross, and V. Greco-Hunt, "Neural networks in radiologic diagnosis—I: Introduction and illustration," *Investigat. Radiol.*, vol. 25, pp. 1012–1016, 1990.

[2] Y. Wu, K. Doi, C. E. Metz, N. Asada, and M. L. Giger, "Simulation studies of data classification by artificial neural networks: Potential applications in medical imaging and decision making," *J. Digit. Imag.*, vol. 6, pp. 117–125, 1993.

[3] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," *Radiology*, vol. 187, pp. 81–87, 1993.

[4] M. A. Kupinski, D. C. Edwards, M. L. Giger, and C. E. Metz, "Ideal observer approximation using Bayesian classification neural networks," *IEEE Trans. Med. Imag.*, vol. 20, pp. 886–899, Sept. 2001.

[5] J. A. Baker, P. J. Kornguth, J. Y. Lo, and C. E. J. Floyd, "Artificial neural network: Improving the quality of breast biopsy recommendations," *Radiology*, vol. 198, pp. 131–135, 1996.

[6] H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Lam, D. D. Adler, and M. M. Goodsitt, "Computerized classification of malignant and benign microcalcifications on mammograms: Texture analysis using an artificial neural network," *Phys. Med. Biol.*, vol. 42, pp. 549–567, 1997.

[7] Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi, "Improving breast cancer diagnosis with computer-aided diagnosis," *Acad. Radiol.*, vol. 6, pp. 22–33, 1999.

[8] Z. Huo, M. L. Giger, C. J. Vyborny, and C. E. Metz, "Breast cancer: Effectiveness of computer-aided diagnosis observer study with independent database of mammograms," *Radiology*, vol. 224, pp. 560–568, 2002.

[9] K. Ashizawa, H. MacMahon, T. Ishida, K. Nakamura, C. J. Vyborny, S. Katsuragawa, and K. Doi, "Effect of an artificial neural network on radiologists' performance in the differential diagnosis of interstitial lung disease using chest radiographs," *Amer. J. Roentgenol.*, vol. 172, pp. 1311–1315, 1999.

[10] P. A. Lachenbruch and M. R. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, pp. 1–11, 1968.

[11] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. Boston, MA: Academic, 1990.

[12] H. P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers," *Med. Phys.*, vol. 26, pp. 2654–2668, 1999.

[13] W. S. Sarle, "Stopped training and other remedies for overfitting," in *Proc. 27th Symp. Interface*, 1995, pp. 352–360.

[14] C. E. Metz, "ROC methodology in radiologic imaging," *Investigat. Radiol.*, vol. 21, pp. 720–733, 1986.

[15] J. A. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 240, pp. 1285–1293, 1988.

[16] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29–36, 1982.

[17] C. E. Metz, "Some practical issues of experimental design and data analysis in radiological ROC studies," *Investigat. Radiol.*, vol. 24, pp. 234–245, 1989.

[18] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in Microstructure of Cognition. Volume 1: Foundations*, D. E. Rumelhart, J. L. McClell, and T. P. R. Group, Eds. Cambridge, MA: The MIT Press, 1986, vol. 1, Computational model of cognition and perception, pp. 318–362.

[19] Y. Jiang, R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vyborny, and K. Doi, "Malignant and benign clustered microcalcifications: Automated feature analysis and classification," *Radiology*, vol. 198, pp. 671–678, 1996.

[20] R. Tibshirani, "A comparison of some error estimates for neural network models," *Neural Comput.*, vol. 8, pp. 152–163, 1996.

[21] G. Papadopoulos, P. J. Edwards, and A. F. Murray, "Confidence estimation methods for neural networks: A practical comparison," *IEEE Trans. Neural Networks*, vol. 12, pp. 1278–1287, Nov. 2001.

The use of *a priori* information in the detection of mammographic microcalcifications to improve their classification

María F. Salfity[a)]
Instituto de Física Rosario, CONICET-UNR, Bv. 27 de Febrero 210 bis, 2000 Rosario, Argentina

Robert M. Nishikawa, Yulei Jiang, and John Papaioannou
Kurt Rossmann Laboratories for Radiologic Image Research, Department of Radiology, MC2026, The University of Chicago, 5841 South Maryland Avenue, Chicago, Illinois 60637

(Received 30 July 2002; accepted for publication 10 January 2003; published 22 April 2003)

In this work, we present a calcification-detection scheme that automatically localizes calcifications in a previously detected cluster in order to generate the input for a cluster-classification scheme developed in the past. The calcification-detection scheme makes use of three pieces of *a priori* information: the location of the center of the cluster, the size of the cluster, and the approximate number of calcifications in the cluster. This information can be obtained either automatically from a cluster-detection scheme or manually by a radiologist. It is used to analyze only the portion of the mammogram that contains a cluster and to identify the individual calcifications more accurately, after enhancing them by means of a "Difference of Gaussians" filter. Classification performances (patient-based A_z=0.92; cluster-based A_z=0.72) comparable to those obtained by using manually-identified calcifications (patient-based A_z=0.92; cluster-based A_z=0.82) can be achieved. © *2003 American Association of Physicists in Medicine.* [DOI: 10.1118/1.1559884]

Key words: CAD, mammography, microcalcifications, detection, classification

I. INTRODUCTION

Screening mammography is the best available tool for detecting cancerous lesions before clinical symptoms appear, and it has been shown to reduce breast cancer mortality.[1,2] Since about half of the cancers detected by mammography correspond to clustered microcalcifications, these lesions are one of the early signs of early breast cancer.[3] However, because of the small size of microcalcifications and the sometimes very slight differences in the appearance of benign and malignant clusters, the differentiation of benign and malignant lesions represents a very complex problem. In fact, it has been reported that only 10%–35% of breast biopsies yields cancer.[4,5]

Computer-aided diagnosis[6] (CAD) can potentially help radiologists improve the diagnosis of malignant and benign breast lesions and as a consequence reduce the number of biopsies performed on benign lesions.[7–10] Several researchers have shown that the radiologists' performance in distinguishing benign from malignant calcifications is statistically significantly improved when they use a computer aid.[11–13]

Researchers at the University of Chicago have developed a computerized method for the classification of clustered microcalcifications.[8,12] Eight features, related to microcalcification size, shape, quantity, and spatial distribution, are automatically extracted from the image. An artificial neural network (ANN) combines these features to produce an estimate of the likelihood of malignancy of each cluster present in the image. This likelihood can then be used by a radiologist as a second opinion to decide whether the microcalcification cluster is malignant or benign. The feature extraction process of this classification method requires as input the x and y locations of each microcalcification. In the previous studies, the locations of the microcalcifications were determined manually. Localizing each calcification in a manual fashion is a time-consuming task and would not be practical for a clinical implementation, considering that the number of calcifications in a cluster can be 100 or even higher. Therefore, the automatic identification of the calcifications prior to the classification of clusters is desired.

Researchers at the University of Chicago have also developed a cluster-detection scheme.[14–20] In order to determine the presence of a cluster in a mammogram, it is not necessary to identify all calcifications. In fact, the average number of calcifications detected by the cluster-detection scheme is about 40%, plus 20% of false-positives.[21] However, the number of calcifications that are identified in a cluster is relevant for classification purposes. Features such as the number of calcifications, the cluster size, and the mean calcification area, are used to distinguish benign from malignant clusters, and their values will depend upon the accuracy of the detection of individual microcalcifications. For these reasons, the cluster-detection and the cluster-classification schemes have not yet been merged into a single unit.

Jiang *et al.*[21] studied the dependence of the ANN classification scheme on the correct detection of individual calcifications. They found that if the average number of calcifications input to the classifier is above 40% of the actual calcifications, plus an average fraction of false signals of below 50%, the performance of the network does not vary in a significant way when compared to the performance of five radiologists. Also, training the ANN with computer-detected microcalcifications degraded the performance of the classification scheme.

In this work, we present a scheme for a more precise

TABLE I. Parameters of the calcification-detection scheme.

Cluster class	Number of true calcifications, N	Range R_c for global threshold	Minimum number of signals S_{min}, for local threshold
1	$N<6$	[30, 50]	3
2	$6\leq N\leq 10$	[50, 100]	6
3	$N>10$	[100, 200]	11

localization of the calcifications once a cluster is detected, in order to generate, automatically, the input for the cluster-classification scheme. The new scheme will be referred to as the calcification-detection scheme to differentiate it from the cluster-detection scheme. We compare the performance of the classification scheme when its input, i.e., the locations of the individual calcifications in the cluster, is provided (a) manually, (b) automatically by the cluster-detection scheme (both cluster and individual calcifications are computer detected in this case), and (c) automatically by different operating points of the calcification-detection scheme presented in this work (the clusters are manually identified in this case).

II. MATERIALS AND METHODS

A. Databases and regions of interest

Two independent mammogram databases were used in this study. All films contained at least one cluster of microcalcifications, the biopsy proven to be either benign or malignant.

Database I consisted of 100 mammograms from 53 patients. Thirty four patients presented benign microcalcification clusters while the remaining 19 had malignant clusters. On average each patient presented two microcalcification clusters, which were either different clusters or the same cluster imaged in different views. Jiang *et al.*[8] reported the performance of five radiologists in the rating of malignancy potential of the clustered microcalcifications, which shows that a large number of cases are difficult to diagnose. The mammograms were digitized with a Fuji drum scanner with a gray-scale resolution of 10 bits and a pixel size and sampling rate of 0.1 mm/pixel. There were a total of 107 clusters (40 malignant, 67 benign), of which 10 belonged to class 1, 39 to class 2, and 58 to class 3. The cluster classes are defined according to the number of calcifications present in the cluster in the first two columns of Table I.

Database II consisted of 237 mammograms from 131 patients. Sixty-six patients presented benign microcalcification clusters while the remaining 65 had malignant clusters. On average, there were 1.8 microcalcification clusters per patient. The films were digitized with a Lumiscan-100 (Lumisys, Sunnyvale, CA) scanner with the same spatial and gray-scale resolution and sampling rate as for database I. Two hundred forty six microcalcification clusters (123 malignant, 123 benign) were present in this set of images. Six clusters belonged to class 1, 62 to class 2, and 178 to class 3.

For each cluster in both databases, a researcher manually identified the microcalcifications by using a high-quality computer monitor and by referencing the film mammograms. For each cluster, its bounding box, the smallest rectangle that contains the entire cluster, was determined by using the manually-identified calcifications. Next, a region of interest defined by the bounding box plus a 55-pixel margin surrounding it was extracted. The additional margin was needed for calculating the features as input to the ANN cluster classifier.

We used database I to determine the appropriate parameters of the calcification-detection scheme, and database II to evaluate the performance of the resulting scheme.

B. Description of the calcification-detection scheme

A flow-chart of the calcification-detection scheme is shown in Fig. 1. The scheme requires two inputs: a region of interest that contains a cluster of microcalcifications and the class to which the cluster belongs according to its number of calcifications, N. Three classes were used:[22] class 1 if $N<6$, class 2 if $6\leq N\leq 10$, and class 3 if $N>10$. This information was used to devise a more accurate calcification segmentation procedure.

The calcifications were first enhanced by means of a Difference of Gaussians (DoG) filter[23] and then segmented via global and local thresholdings. The DoG filter smooths the image with two Gaussian kernels of different standard deviations, σ_1 and σ_2, and then subtracts one smoothed version of the image from the other. Database I was used to empirically select the values of $\sigma_1=1.1$, $\sigma_2=1.4$ with kernel sizes of 7×7 and 9×9 pixels, respectively. With these parameters, the effect of the filter was to enhance signals of the typical microcalcification of size 3×3 pixels.

A global and a local thresholding operations then segmented the enhanced potential calcifications (referred to as signals in this paper). The global thresholding kept a number of signals within the range, R_c, where R_c depended on the cluster class (Table I). The minimum and maximum boundaries of R_c were empirically set to lie well above the expected number of calcifications, N, in order to increase the chances of thresholding all the actual calcifications. The local thresholding was applied to the kept signals in order to reduce false-positives. In this step, a minimum number of signals, S_{min}, was always segmented where S_{min} depended on the cluster class and was set as the minimum number of calcifications required for the cluster to belong to each class (Table I).

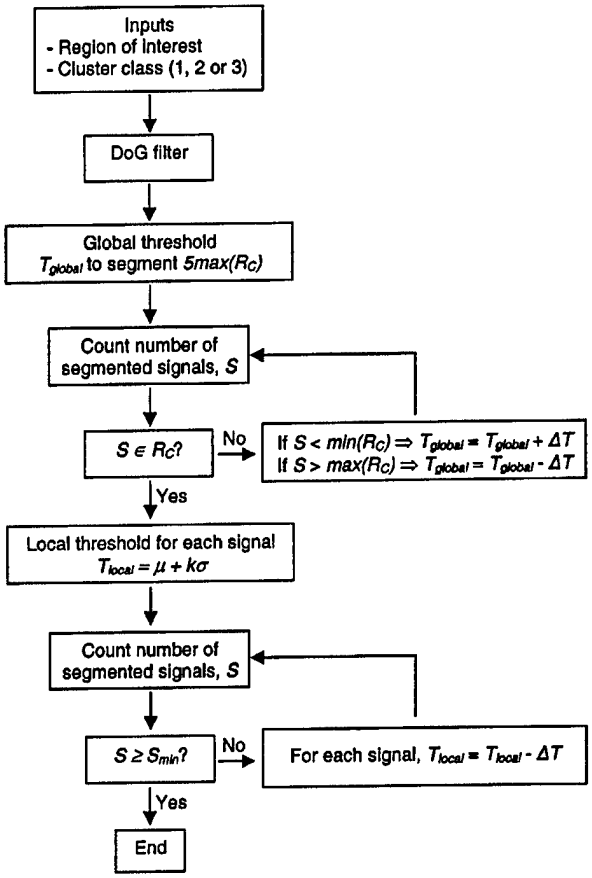


FIG. 1. Flow-chart of the calcification-detection scheme.

The global thresholding was iterative. The initial global threshold T_{global} was set as the gray level such that the number of pixels above it equaled to five times the upper limit of R_c. Here it was assumed that the average signal size is five pixels, which agrees with typical calcification areas. After applying this threshold, the number of candidate calcifications, S, was counted. Signals with an area of 1 pixel or larger than 100 pixels were excluded because they are not likely to be real microcalcifications. Also, signals that were within 55 pixels from the edge of the region of interest were ignored because this area was not part of the cluster bounding box. If S did not lie within R_c, T_{global} was increased or decreased by a small step, ΔT, where ΔT was selected empirically as 0.1. Note that the gray level of the image was no longer quantized as in the original image after the DoG filtering. This process was repeated until the number of signals S fell within R_c.

The local threshold was also iterative. Centered on each signal identified in the previous step, a 100×100-pixels box was defined. In this box, the mean μ and standard deviation σ of the background gray levels were calculated by excluding those pixels that were identified in the global threshold step as potential signals. The initial local threshold T_{local} was set as $\mu + k\sigma$, where k is a variable parameter. If the maximum gray level of the signal was below T_{local}, the signal was discarded. Once all signals had been analyzed in this way, the number of remaining signals S was compared to S_{min} (Table I). If $S < S_{min}$, T_{local} was decreased by ΔT. Again, S was calculated and compared to S_{min}. This process was repeated until $S \geq S_{min}$.

The output of the calcification-detection scheme is the center-of-mass coordinates of detected signals and can be used as input to the classification scheme.

C. Evaluation of the performance of computer schemes for the detection of individual calcifications

The performance of the computerized detection of individual calcifications was evaluated by counting the number of signals that matched actual calcifications (true-positive signals) and the number of signals that did not have such correspondence (false-positive signals). A signal was considered a true-positive if its center of mass lay within five pixels from a true calcification. True-positive detection, TPD_s, can be defined as the ratio of the number of true-positive signals to the total number of calcifications present in the cluster.[21] In the same way, false-positive detection, FPD_s, can be defined as the ratio of the number of false-positive signals to the total number of calcifications present in the cluster. Therefore, for manually-identified calcifications $TPD_s = 100\%$ and $FPD_s = 0\%$. Note that FPD_s can be larger than 100%.

D. Computerized classification of clustered microcalcifications

1. Description of the classification scheme

The classification scheme required as input the x and y locations of the microcalcifications. First, the microcalcifications were segmented and eight features that describe calcifications both individually and as a cluster, were automatically extracted. The features, described in detail in Ref. 8, include (a) the number of microcalcifications in a cluster, (b) the mean area, (c) the mean effective volume, (d) the relative standard deviation of the effective thickness, (e) the relative standard deviation of the effective volume, (f) the second highest shape irregularity value, (g) the cluster area, and (h) the cluster circularity. These features were fed to a feed-forward ANN with one hidden layer of six units and an output layer of one unit. The ANN output was related to the likelihood of malignancy of the cluster.

2. ANN training

ANN training was performed with the error-back-propagation algorithm in a leave-one-patient-out fashion. All clusters that corresponded to one patient were set aside as a test set, and the remaining clusters were used for training. This procedure was then repeated for the next patient, until all clusters were classified.

3. Different input data to the classification scheme

In this work, we compare the performance of the cluster classifier when its input was provided (a) manually identify-

ing the calcifications, (b) automatically identifying both the cluster and the calcifications (with the cluster-detection scheme), and (c) automatically identifying the calcifications (with the calcification-detection scheme) in manually identified clusters. The ANN was re-trained for the different input data, i.e., for the features that were extracted when the x and y locations of the individual calcifications were given by (a), (b), and (c). Note that (c) also included several data sets, each corresponding to a different operating point of the calcification-detection scheme.

E. Evaluation of the performance of the classification scheme

The performance of the classifier was evaluated in two different ways: per patient and per cluster.[8] The per cluster analysis was direct because each cluster was given a malignancy rating by the ANN. However, two or more different clusters or the same cluster imaged in different views, may be from the same patient. A radiologist would analyze all clusters in all available views in order to diagnose a patient. The approach taken for the per patient analysis was to keep only the maximum malignancy rating of all clusters associated with the same patient. Receiver operating characteristic (ROC) analysis[24,25] was used to evaluate the performance of the classifier on both the per patient and the per cluster bases. ROC curves, as well as the area A_z and partial area[26] ${}_{0.90}A_z$ under the curves were estimated with Metz's LABROC4 software.

III. RESULTS

A. Detection of individual calcifications

The performance of the cluster-detection scheme for the detection of individual calcifications, using database I, was: TPD_s=55%±21% and FPD_s=20%±33%. These numbers were obtained by analyzing only true-positive detected clusters. The sensitivity in terms of cluster detection for this database was 78% with 2.5 false positive clusters per image.

A more accurate automatic identification of individual calcifications was achieved by the calcification-detection scheme (Fig. 2).

B. Performance of the classification scheme for the different input data

1. Manually-identified calcifications

When manually-identified calcifications from database I were used as the classifier input, the $A_z$value equaled 0.92 on a per patient analysis and 0.83 on a per cluster analysis.[8] The corresponding partial area indices were 0.82 and 0.48 (Table II).

2. Automatically-identified clusters and calcifications with the cluster-detection scheme

The performance of the classifier was substantially degraded when the microcalcifications detected by the cluster-detection scheme were used as the classifier input. The area under the ROC curve had a value of 0.81 on a per patient analysis and 0.79 on a per cluster analysis. The corresponding partial area indices were 0.33 and 0.29 (Table II).

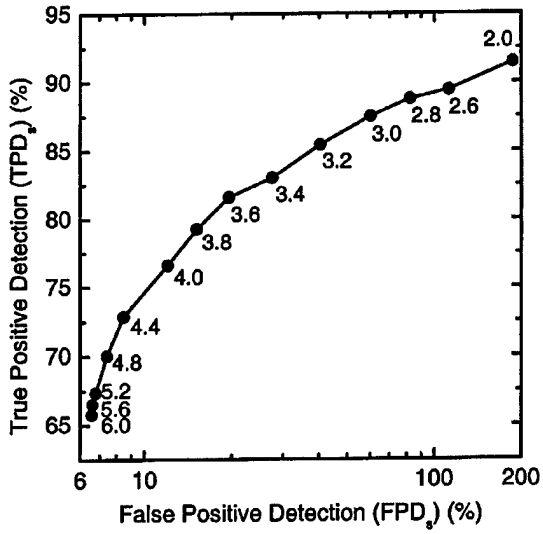


FIG. 2. FROC curve of the calcification-detection scheme on database I. The numbers indicate the value of k associated with each point.

3. Automatically-identified calcifications with the calcification-detection scheme

Figure 3 shows the A_z and ${}_{0.90}A_z$ of the ANN as a function of the calcification-detection parameter, k, in both the per patient (a), (b) and per cluster (c), (d) analyses. The straight solid and dashed lines represent the index values obtained when the classification scheme input was provided by manually-identified calcifications and by the cluster-detection scheme, respectively (Table II). The ROC curves of these two cases are compared to the curve obtained when k=3.2 in Fig. 4. Table III compares the area and partial area indices shown in Fig. 3 to the results of manual identifications. The fractions of cancers correctly classified (on a per patient basis) at fixed false positive fractions of 50% and 30% are shown in Table IV.

C. Evaluation on an independent mammogram database

The calcification-detection scheme with k=3.4, was run on database II. The calcification-detection scheme identified 74±16% of the actual calcifications and 34±48% false positive detections. When these computer-detected signals were used as input to the cluster classifier, the area index equaled 0.89 on a per patient analysis and 0.93 on a per cluster analysis, and the corresponding partial area indices were 0.28 and 0.51. These results are compared to the values obtained when manually-identified signals were used as the classifier input in Table V.

IV. DISCUSSION

The calcification-detection scheme presented in this work had higher performance than the cluster-detection scheme in

TABLE II. The area A_z and partial area ${}_{0.90}A_z$ under the ROC curves obtained on a per patient and on a per cluster basis, on database I, when the input of the classification scheme is given by manual identifications and by the cluster-detection scheme (± indicate standard deviations).

	Per patient		Per cluster	
Classifier input	A_z	${}_{0.90}A_z$	A_z	${}_{0.90}A_z$
Manually	0.92 ± 0.04	0.82 ± 0.08	0.83 ± 0.04	0.48 ± 0.09
Cluster-detection	0.81 ± 0.06	0.33 ± 0.16	0.79 ± 0.05	0.29 ± 0.10

the task of automatically identifying individual calcifications. Compared to the performance of the latter scheme (TPD_s=55% and FPD_s=20%), the proposed calcification detection scheme achieved TPD_s values between 81% and 66% while FPD_s remained below 20%, when $3.6 \leq k \leq 6.0$ (Fig. 2). For k above 5.2 the change in performance was only slight, and the TPD_s and FPD_s values remained around 66% and 7%, respectively. This was a consequence of the local thresholding that guaranteed the detection of a minimum number of signals S_{min}, regardless of the magnitude of k. The detection performance achieved on the independent database with k=3.4 (TPD_s=74%, FPD_s=34%) was slightly lower than the result obtained with the same k value on database I (Fig. 2).

The detection performance decreased with an increasing number of calcifications. For example, for database I and k =3.2, the average true-positive and false-positive detection rates were 95% and 14%, respectively, for class 1 clusters, 86% and 24%, respectively, for class 2 clusters, and 83% and 55%, respectively, for class 3 clusters. For database II and k=3.4, the true-positive detection rate was 100% and the false-positive detection rate was 3% for class 1 clusters, while the respective rates were 79% and 17% for class 2 clusters, and 72% and 40% for class 3 clusters. These different performances are related to the global thresholding parameters as explained later.

It should be noted that the cluster-detection scheme is optimized for the detection of clusters and not of individual calcifications, hence the reason for the implementation of an intermediate step to localize individual calcifications more accurately with the calcification-detection scheme. This more accurate localization is essential in the feature extraction process that generates the input to the ANN so as not to degrade the classifier performance. In fact, the results indicate that high fractions of false-positive signals or low fractions of true-positive signals produce lower performance than those

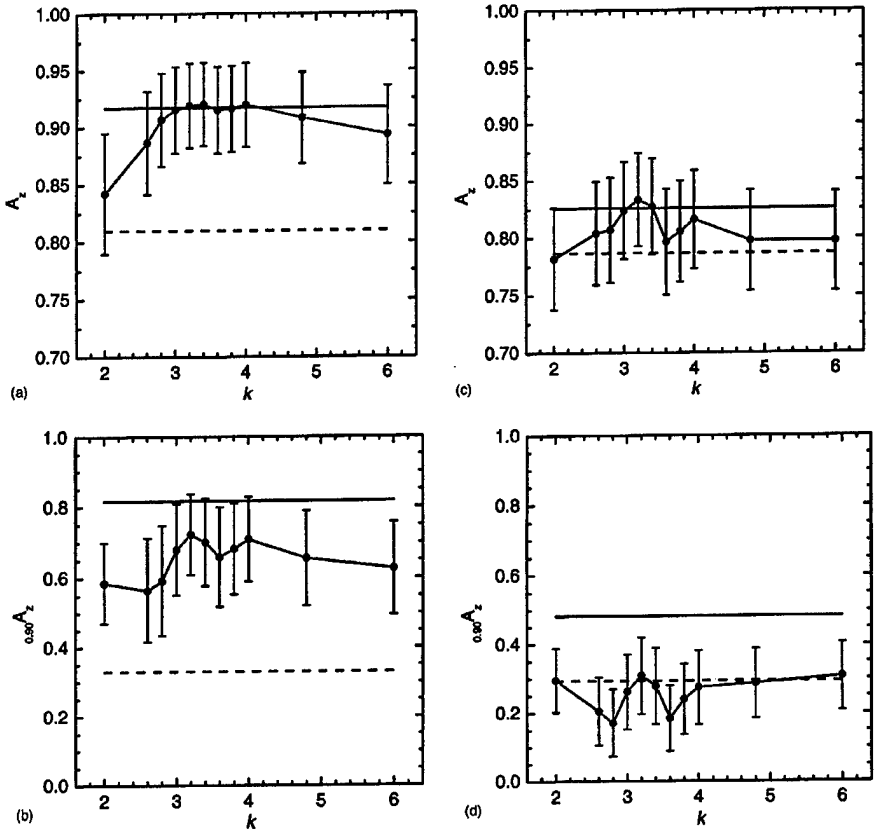


FIG. 3. The area A_z and partial area ${}_{0.90}A_z$ indices obtained on database I when the classifier input is provided by the calcification-detection scheme are shown as a function of the threshold parameter k. The A_z and ${}_{0.90}A_z$ values obtained when manual identifications (—) and when the cluster-detection scheme (---) are used instead, are included for comparison. (a) A_z, per patient analysis. (b) ${}_{0.90}A_z$, per patient analysis. (c)A_z, per cluster analysis. (d) ${}_{0.90}A_z$, per cluster analysis.

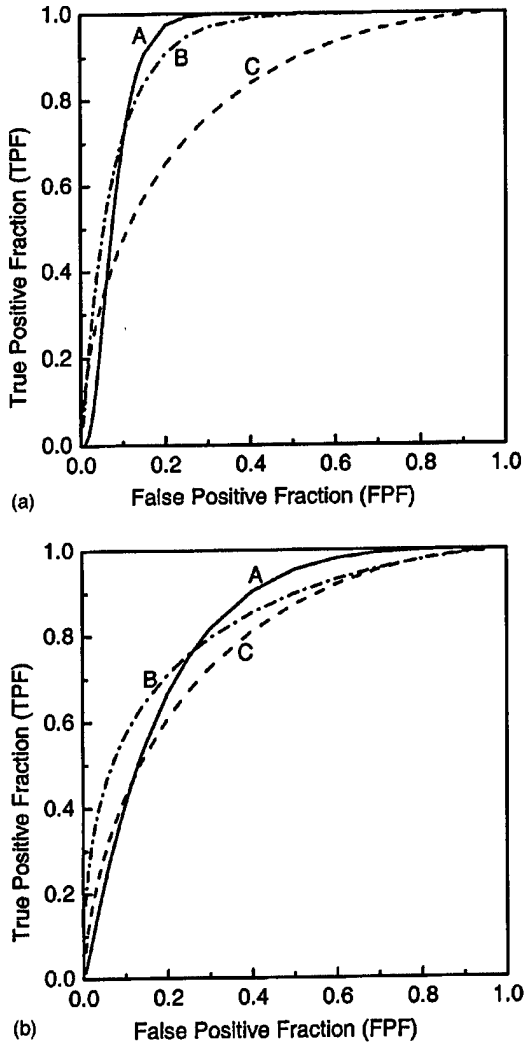


FIG. 4. ROC curves obtained on database I (a) on a per patient basis and (b) on a per cluster basis, when the input of the classification scheme is given by manual identifications (curve A), by the calcification-detection scheme with k=3.2 (curve B), and by the cluster-detection scheme (curve C).

obtained with manually marked calcifications. In particular, when the cluster-detection scheme provided the input of the classification scheme, the performance of the latter was lower than that obtained when manual detections were used (Table II). This degradation in performance was due to the incorrect detection of individual calcifications -55% TPD_s and 20% FPD_s—that affected the features used by the ANN to classify the clusters.

When the calcification-detection scheme was used to identify the calcifications, the classification performance depended on the operation point of the detection scheme (Fig. 3). For $3.0 \leqslant k \leqslant 4.0$, i.e., for ($TPD_s$, FPD_s) pairs between (0.87, 0.60) and (0.77, 0.12), the classifier performance did not substantially change. This agreed with Ref. 21, where it is reported, for the same database, that when using composite computer-detected calcifications the classifier performance remained approximately constant for $TPD_s > 40\%$ ($FPD_s = 0\%$) and for $FPD_s < 50\%$ ($TPD_s = 42\%$). It should be noted that in that study the false-positive detection values were always below 50%, as opposed to this work where a wider range of values was analyzed. Higher or lower values of k resulted in poorer performance of the classifier. This is clearer for $k < 3.0$, when FPD_s increased more rapidly, than for $k > 4.0$, when the calcification-detection performance varied more moderately with low FPD_s values.

It should be noted that in Figs. 3(a) and 3(c) for $3.0 \leqslant k \leqslant 4.0$, both patient- and cluster-based A_z values are close to the respective values obtained by using manual identifications, and higher than the results obtained from the cluster-detection scheme. In the same range of k, the per patient partial area indices ${}_{0.90}A_z$ [Fig. 3(b)] were less than the corresponding value of manual identifications but well above the value obtained from the cluster-detection scheme. However, the cluster-based ${}_{0.90}A_z$ [Fig. 3(d)] was close to the value obtained from the cluster-detection scheme.

In fact, the patient- and cluster-based area indices A_z values were not significantly different from the values obtained with manual identifications (Table III). However, as can be appreciated in Fig. 4, the ROC curves obtained by using the calcification-detection scheme crossed the ROC curves of manual identifications in such a way that the A_z indices did not substantially differ but the ${}_{0.90}A_z$ values did. The partial area index ${}_{0.90}A_z$ was not significantly different from the value corresponding to manual identifications only for k=3.2, 3.4, 3.6, and 4.0 in the per patient analysis. These k values correspond to detection performances, expressed as (TPD_s, FPD_s) pairs, of: (0.85, 0.40), (0.83, 0.27), (0.82, 0.20), and (0.77, 0.12), respectively. In the per cluster analysis, the partial area values were always significantly different from the manual identification value. Similar results were obtained on the independent database II (Table V). The patient- and cluster-based A_z indices did not differ significantly from the values obtained with manual identifications, and the partial area index ${}_{0.90}A_z$ was not significantly different in the per cluster analysis.

The differences between the per patient and per cluster performances observed in Fig. 3 arise as a consequence of keeping the maximum malignancy rating of all clusters associated with the same patient in the per patient analysis (Sec. II E). When there is more than one cluster per patient, this can result in an equal or improved classification performance in malignant cases, and in an equal or worse performance in benign cases. For instance, in database I, where there is an average of 2 clusters per patient, the differences between the per patient and per cluster classification results when manually-identified calcifications were used (Table II), can be explained by analyzing the average ANN outputs for benign and malignant cases. In the per cluster analysis the average ANN output was 0.66 for malignant clusters and 0.30 for benign ones, while in the per patient analysis the respective values were 0.78 and 0.38. The standard deviations of the ANN outputs in both benign and malignant cases were comparable in the per cluster and per patient analyses. Therefore, in the per patient analysis there was less overlap

TABLE III. The area A_z and partial area $_{0.90}A_z$ indices obtained on database I (a) by using manually-identified calcifications and (b) by using the calcification-detection scheme. p-values were calculated with CLABROC.

	A_z			$_{0.90}A_z$		
k	(a)	(b)	p-value	(a)	(b)	p-value
Per patient						
2.0	0.92	0.84	0.077	0.82	0.58	0.005
2.6	0.92	0.89	0.233	0.82	0.56	0.005
3.0	0.92	0.91	0.558	0.82	0.68	0.013
3.2	0.92	0.92	0.907	0.82	0.72	0.052
3.4	0.92	0.92	0.774	0.82	0.70	0.066
3.6	0.92	0.91	0.815	0.82	0.66	0.054
3.8	0.92	0.92	0.608	0.82	0.68	0.043
4.0	0.92	0.92	0.855	0.82	0.71	0.077
4.8	0.92	0.91	0.451	0.82	0.65	0.011
6.0	0.92	0.89	0.305	0.82	0.63	0.009
Per cluster						
2.0	0.82	0.78	0.121	0.48	0.29	0.002
2.6	0.82	0.80	0.577	0.48	0.20	<0.001
3.0	0.82	0.83	0.636	0.48	0.26	0.001
3.2	0.82	0.83	0.979	0.48	0.31	0.005
3.4	0.82	0.83	0.920	0.48	0.28	0.003
3.6	0.82	0.80	0.248	0.48	0.18	<0.001
3.8	0.82	0.80	0.286	0.48	0.24	0.001
4.0	0.82	0.81	0.430	0.48	0.27	0.003
4.8	0.82	0.80	0.186	0.48	0.28	0.030
6.0	0.82	0.80	0.155	0.48	0.31	0.016

between the two distributions than in the per cluster analysis, and as a consequence the classification performance improved. In database II, where there is an average of 1.8 clusters per patient, when manually-identified calcifications were used (Table V), there was practically no difference between the per patient and per cluster A_z and $_{0.90}A_z$ indices. In this case, the average ANN output for benign cases increased from 0.31 (per cluster) to 0.36 (per patient), and the average ANN output for malignant cases increased from 0.66 (per cluster) to 0.71 (per patient). Therefore the difference between the means of the two distributions did not change between the per cluster and per patient analyses and as a consequence there was no difference in performance.

The fact that the partial area indices were consistently lower for computer identifications than for manual ones, could be explained by analyzing the calcification-detection scheme performance for benign and malignant clusters separately (Fig. 5). At each k value, FPD_s are higher for malignant than for benign clusters. For $k \leq 4.0$, TPD_s are similar for benign and malignant cases, but for $k > 4.0$, TPD_s are lower for malignant than for benign clusters. This means that the detection performance was lower in malignant than in benign cases. A similar trend was observed on the results obtained with database II (Sec. III C), for which $TPD_s = 80\%$ and $FPD_s = 36\%$ in benign cases as opposed to $TPD_s = 69\%$ and $FPD_s = 31\%$ on malignant cases. This difference in performance is related to two factors. First, malignant clusters in general contain more calcifications than benign ones. In particular, for database I, the mean number of calcifications was 28 for malignant clusters and 10 for benign ones, while for database II the respective mean values were 34 and 16. Second, the global thresholding kept a disproportionately larger number of signals when the estimated number of calcifications, N, was large (Table I). This is reflected in the different detection performances across the cluster classes as noted previously. In database I, clusters with more than 10 calcifications represented 92% of malignant clusters as opposed to 31% of benign clusters, while in database II the respective fractions were 83% and 61%. Therefore, malignant clusters tended to yield both a higher number of false-positive signals and a higher FPD_s value, particularly for $k \leq 4.0$. For $k > 4.0$, the difference in FPD_s values in malignant and benign clus-

TABLE IV. A comparison between the sensitivity levels achieved by the ANN cluster classifier (per patient analysis), on database I, at 30% and 50% false positive fractions, when the classifier input is provided by manually-identified calcifications and by the calcification-detection scheme with $k=3.2$. The numbers in parenthesis indicate 95% confidence intervals. p-values were calculated with CLABROC.

	Manual identifications	Calcification-detection scheme $k=3.2$	p-value
TPF (%) at 30% FPF	99.87 (71.11, 100)	97.10 (74.56, 99.91)	0.12
TPF (%) at 50% FPF	100 (86.05, 100)	99.75 (85.59,100)	0.09

TABLE V. The area A_z and partial area $_{0.90}A_z$ obtained on database II, by using manually-identified calcifications and by using the calcification-detection scheme with k=3.4.

A_z			$_{0.90}A_z$		
Manual identification	Calc-detection scheme	p-value	Manual identification	Calc-detection scheme	p-value
Per patient					
0.91 ± 0.02	0.89 ± 0.03	0.294	0.62 ± 0.08	0.28 ± 0.13	0.003
Per cluster					
0.93 ± 0.02	0.93 ± 0.02	0.982	0.62 ± 0.07	0.51 ± 0.10	0.436

ters was smaller than for $k \leq 4.0$, but TPD_s in malignant clusters were lower than in benign ones. As a consequence, classification results were more degraded for malignant than for benign clusters, when compared to results obtained from manual identifications. This is reflected in the reduced partial area index (Tables III and IV). Therefore, by eliminating false-positive detections the classifier performance in the area of higher sensitivity could be improved further. This could be achieved by redefining the third cluster class (Table I) and adding a fourth class to avoid detecting an excessive number of signals when $N>10$. Further improvement could be obtained by introducing to the calcification-detection scheme a more efficient false-positive reduction step, such as a neural network that takes as input several signal features.[27]

In spite of the lower partial area indices, for database I, at k=3.2, 99.75% cancers were correctly classified on a per patient basis, while 50% of benign cases were classified as malignant. Alternatively, at a false-positive fraction of 30%, a sensitivity of 97.1% was achieved. These sensitivity values are not significantly different from those obtained by using manual identifications (Table IV).

A fully automated system could be realized by linking the cluster and calcification detection schemes, i.e., to identify the cluster with the cluster-detection scheme and then to identify the calcifications in that cluster with the calcification-detection scheme. The cluster location and size, and the number of calcifications in the cluster will be provided by the cluster-detection scheme. However, since the cluster-detection scheme is not optimized for the detection of individual calcifications, this information might not be correct. This needs to be taken into account and compensated for if the computer schemes are to be linked. The implementation of a fully-automated system for the detection and classification of clustered microcalcifications is part of our future work.

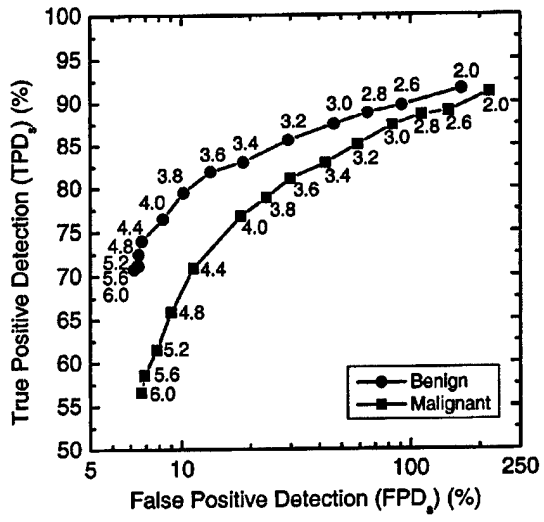


FIG. 5. FROC curves of the calcification-detection scheme for benign and malignant clusters (database I). The numbers indicate the value of k associated with each point.

Ideally, a fully automated system would be best suited for clinical purposes. However, automated detection schemes do not have 100% sensitivity. Furthermore, commercial systems are designed to detect cancer, not necessarily suspicious benign lesions. Therefore, there will be calcification clusters that are not detected by the computer, but that look suspicious to the radiologists. In those situations an interface is needed so that the radiologist can indicate to the computer a suspicious region to analyze. We believe that our interface is less burdensome than requiring the radiologist to identify all calcifications in the cluster and it produces more accurate results than segmenting the ROI without the approximate number of calcifications in the cluster. We do not feel that it will be difficult for a radiologist to indicate one of four different categories for the number of calcifications in the cluster. The radiologist does not have to count the exact number of calcifications, since the categories are fairly broad. We are currently conducting a study to determine the effect on the classifier's performance of inaccurately assigning a category.

Since the accuracy of classification of a mass or mass-like lesion depends on the accuracy of the segmented lesion, improvements in the segmentation of the lesion should improve the classification result.[28] *A priori* information could be used to improve the segmentation. For example, the approximate size and information on shape (round, oval, lobular, or irregular), margin (circumscribed, microlobulated, obscured, indistinct, or spiculated), as well as the local breast density could be used to improve the segmentation, in a similar fashion to what we have described in this paper for microcalcifications. The *a priori* information could come from a radiologist, for example, using a softcopy interface or from an initial detection scheme.

In conclusion, we have shown that by using *a priori* in-

formation about the cluster, a higher percentage of individual microcalcifications can be identified. By doing so, classification of the cluster (benign versus malignant) can be done more accurately and at a level comparable to the result if all the microcalcifications were identified manually.

ACKNOWLEDGMENTS

The authors wish to thank Charles E. Metz for the use of his LABROC4 and CLABROC programs for ROC analysis. Funding was provided in part by a grant from the National Cancer Institute (CA 60187). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of any of the supporting organizations. María Fernanda Salfity is supported by a scholarship from Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina. Robert M. Nishikawa, and John Papaioannou are shareholders in R2 Technology, Inc. (Los Altos, CA). It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interests which may appear to be affected by the research activities.

[a]Electronic mail: m.f.salfity@lboro.ac.uk

[1] S. A. Feig, "Decreased breast cancer mortality through mammographic screening: Results of clinical trials," Radiology **167**, 659–665 (1988).

[2] C. R. Smart, R. E. Hendrick, J. H. Rutledge, and R. A. Smith, "Benefit of mammography screening in women ages 40 to 49 years: Current evidence from randomized controlled trials," Cancer **5**, 1619–1626 (1995).

[3] CancerNet 2000, a service of the National Cancer Institute, http://www.cancernet.nci.nih.gov.

[4] E. A. Sickles, "Mammographic features of 300 consecutive nonpalpable breast cancers," Am. Q. Roentgenol. **146**, 661–663 (1986).

[5] A. M. Knutzen and J. J. Gisvold, "Likelihood of malignant disease for various categories of mammographicaly detected, nonpalpable breast lesions," Mayo Clin. Proc. **68**, 454–460 (1993).

[6] K. Doi, M. L. Giger, R. M. Nishikawa, K. R. Hoffmann, H. MacMahon, R. A. Schmidt, and K. G. Chua, "Digital radiography. A useful clinical tool for computer-aided diagnoses by quantitative analysis of radiographic images," Acta Radiol. **34**, 426–439 (1993).

[7] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer," Radiology **187**, 81–87 (1993).

[8] Y. Jiang, R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vyborny, and K. Doi, "Malignant and benign clustered microcalcifications: automated feature analysis and classification," Radiology **198**, 671–678 (1996).

[9] J. A. Baker, P. J. Kornguth, J. Y. Lo, and C. E. J. Floyd, "Artificial neural network: improving the quality of breast biopsy recommendations," Radiology **198**, 131–135 (1996).

[10] H.-P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Lam, D. D. Adler, and M. M. Goodsitt, "Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network," Phys. Med. Biol. **42**, 549–567 (1997).

[11] D. J. Getty, R. M. Pickett, C. J. D'Orsi, and J. A. Swets, "Enhanced interpretation of diagnostic images," Invest. Radiol. **23**, 240–252 (1988).

[12] Y. Jiang, R. M. Nishikawa, R. A. Schimdt, C. E. Metz, M. L. Giger, and K. Doi, "Improving breast cancer diagnosis with computer-aided detection," Acad. Radiol. **6**, 22–33 (1999).

[13] H.-P. Chan, B. Sahiner, M. A. Helvie, N. Petrick, M. A. Roubidoux, T. E. Wilson, D. D. Adler, C. Paramagul, J. S. Newman, and S. Sanjay-Gopal, "Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study," Radiology **212**, 817–827 (1999).

[14] H.-P. Chan, K. Doi, C. J. Vyborny, H. MacMahon, and P. M. Jokich, "Image feature analysis and computer-aided diagnosis in digital radiography. 1. Automated detection of microcalcifications in mammography," Med. Phys. **14**, 538–548 (1987).

[15] H.-P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu, and H. MacMahon, "Improvement in radiologists' detection of clustered microcalcifications on mammograms: The potential of computer-aided diagnosis," Invest. Radiol. **25**, 1102–1110 (1990).

[16] R. M. Nishikawa, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt, "Computer-aided detection of clustered microcalcifications: An improved method for grouping detected signals," Med. Phys. **20**, 1660–1666 (1993).

[17] R. M. Nishikawa, Y. Jiang, M. L. Giger, R. A. Schmidt, C. J. Vyborny, W. Zhang, J. Papaioannou, U. Bick, R. Nagel, and K. Doi, "Performance of automated CAD schemes for the detection and classification of clustered microcalcifications," in *Digital Mammography*, edited by A. G. Gale, S. M. Astley, D. R. Dance, and A. Y. Carins (Elsevier, Amsterdam, Holland, 1994), pp. 13–20.

[18] W. Zhang, K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt, "An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms," Med. Phys. **23**, 595–601 (1996).

[19] R. H. Nagel, R. M. Nishikawa, J. Papaioannou, and K. Doi, "Analysis of methods for reducing false positives in the automated detection of clustered microcalcifications in mammograms," Med. Phys. **25**, 1502–1506 (1998).

[20] R. M. Nishikawa, M. L. Giger, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Prospective testing of a clinical CAD workstation for the detection of breast lesions on mammograms," in *Computer-Aided Diagnosis in Medical Imaging*, edited by K. Doi, H. MacMahon, M. L. Giger and K. R. Hoffmann (Elsevier, Amsterdam, Holland, 1998), pp. 209–214.

[21] Y. Jiang, R. M. Nishikawa, and J. Papaioannou, "Dependence of computer classification of clustered microcalcifications on the correct detection of microcalcifications," Med. Phys. **28**, 1949–1957 (2001).

[22] American College of Radiology, *Breast Imaging Reporting and Data System* (American College of Radiology, Reston VA, 1998).

[23] J. C. Russ, *The Image Processing Handbook* (CRC, Boca Raton, 1992).

[24] C. E. Metz, "ROC methodology in radiologic imaging," Invest. Radiol. **21**, 720–733 (1986).

[25] C. E. Metz, "Some practical issues of experimental design and data analysis in radiological ROC studies," Invest. Radiol. **24**, 234–245 (1989).

[26] Y. Jiang, C. E. Metz, and R. M. Nishikawa, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests," Radiology **201**, 745–750 (1996).

[27] D. C. Edwards, M. A. Kupinski, R. Nagel, R. M. Nishikawa, and J. Papaioannou, "Using a Bayesian Neural Network to optimally eliminate false-positive microcalcification detections in a CAD scheme," in *Digital Mammography*, edited by M. J. Yaffe (Medical Physics Publishing, Toronto, Canada, 2001), pp. 168–173.

[28] Z. Huo and M. L. Giger, "Evaluation of a computer segmentation method based on performances of an automated classification method," Proc. SPIE **3981**, 16–21 (2000).

Experience with Computer-Aided Detection in a Low-Volume Mammography Clinic

Carl J. Vyborny[a,b], Catherine Kukec[b], Yulei Jiang[a] and Kunio Doi[a]

[a)] Department of Radiology, The University of Chicago, Chicago, Illinois
[b)] Department of Radiology, LaGrannge Memorial Hospital, LaGrange, Illinois
cvyborny@chi-rad-soc.org

Abstract. The effect of the R2 ImageChecker computer-aided detection system was evaluated in a low-volume mammography clinic in the suburbs of Chicago. Prompts from the ImageChecker resulted in an additional 1.0% of patients being recalled for additional imaging while increasing the yield for early detection of impalpable breast cancer by 5.2% over a 24 month period. Cancer detection rates before and after installation of the computer-aided detection system were not statistically different.

1. Introduction

It is well documented that present generation computer-aided detection (CAD) devices in mammography can detect breast cancers that are visible retrospectively on mammograms but that were not reported by the original interpreting radiologist [1,2]. The utility of CAD devices has now also been evaluated prospectively in screening centers [3,4]. We report our experience with the R2 ImageChecker at a small imaging clinic doing approximately 10 to 12 mammograms a day.

2. Materials and Methods

The study described here was performed after approval by, and under supervision of, the applicable Institutional Review Board. The R2 ImageChecker was installed at Grant Square Imaging in the suburbs of Chicago on April 1, 1998. A number of conventional mammography practice parameters were monitored between January 1, 1997 and March 31, 2000 in order to assess the effect and utility of the ImageChecker. Additionally, commencing on August 1, 1998, radiologists were asked to record cases in which patients were recalled to the department for additional imaging as the result of CAD prompts.

The volume of mammography at the clinic averaged between 10 and 12 cases per day throughout the course of the study. A single radiologist staffed the site and, in addition to mammograms, also interpreted plain film studies, as well as computed

tomography, nuclear medicine and ultrasound examinations with the total case volume being approximately 40 studies per day. The volume and mix of the examinations was such that radiologists at the clinic were about "half as busy" as at other sites staffed by the radiology professional group. Four radiologists read more than 95% of the mammograms at the clinic during the 39 month study period.

It was the policy of the clinic that no added charges be assessed for additional mammography views. The technologists therefore routinely reviewed the cases before patient discharge and brought potential abnormalities to the radiologist's attention who then decided whether additional views might be needed before the examination was considered complete. This practice style did not change after installation of the ImageChecker, with technologist review coming before the films were digitized for CAD analysis. The approach described tended to reduce significantly the percentage of patients "called back" to the department on a separate date for additional views. It also reduced the potential for observational oversights due to a single reading by one radiologist.

3. Results

Between August 1, 1998 and March 31, 2000, 51 of 5359 (1.0%) patients were called back to the department based on prompts by the ImageChecker. An independent assessment of the average recall rate for the three radiologists who interpreted examinations throughout the entire study period showed an increase from 3.1% between January 1, 1997 and March 31, 1998 (before installation) to 4.2% between the August 1, 1998 and March 31, 2000 ($p=0.037$; Pearson Chi-square test with Yates correction). As described above, the practice style at the clinic tended to reduce recall rates appreciably.

The 51 additional patient recalls resulted in 6 biopsies, one of which yielded a malignant diagnosis. This represented an increase of 5.2% (1/19) in the yield of impalpable cancers at the clinic during the monitored period.

No statistically significant difference in the detection rate for impalpable breast cancer was noted before and after ImageChecker installation. Between January 1, 1997 and March 31, 1998 the detection rate was 4.5/1000 (13/2866). Between April 1, 1998 and March 31, 2000 the detection rate was 3.8/1000 (24/6345), ($p=0.73$). The positive predictive values for biopsy before and after installation were 38% and 30% respectively ($p=0.55$). In cases for which T staging was available, the minimal cancer detection rate (percentage Tis, T1a and T1b lesions) was 50% (4/8) before installation and 75% (9/12) after installation ($p=0.36$; 2-tailed Fisher exact test).

4. Discussion

The practice setting in this study is one that could be reasonably expected to minimize the benefits of computer-aided detection. The radiologists reviewed a relatively small number of cases every day and did so in a rather relaxed environment. Additionally, the technologists prescreened examinations for potential abnormalities in order to reduce recall rates. In so doing, they often served effectively as second readers, an approach taken at some institutions to improve the sensitivity of screening mammography [5]. It is therefore of interest that the ImageChecker alerted radiologists to abnormalities, overlooked on original film review, but warranting patient recall in nearly 1.0% of the cases. Further, biopsy recommendations resulted in more than 10% of the recalled patients (6/51). These results suggest that even in a favorable practice environment, radiologists do not perceive all findings that, retrospectively, can be considered worthy of concern.

The question does arise whether CAD information is as beneficial in a low-volume reading situation. In particular, one could postulate that CAD prompts might make the unhurried radiologist hyperattentive to borderline findings that ultimately prove to be benign. It has been shown that such borderline findings, at least when identified by experts, almost never are the result of malignant lesions [6]. The percentage of patients additionally recalled due to CAD prompts in this study was approximately 1.0%, a value similar to that reported by Freer and Ulissey. At the same time, the incremental increase in yield for breast cancer detection was much less in this study, i.e., 5%, than the nearly 20% reported by Freer and Ulissey in a study carried out in a much busier mammography center. Given the relatively small number of cancer cases involved, however, the difference is not statistically significant ($p=0.27$; 2-tailed Fisher exact test). Radiologists using CAD information in any setting should nevertheless be aware of the very low yield in the recall of borderline findings.

There are obvious difficulties in arriving at statistically meaningful observations in low-case-volume practice. We were fortunate to have a very stable practice environment for the 39 month study period during which slightly more than 9000 examinations were performed. Given the very low incidence of breast cancer, however, our results fail to prove or disprove the benefit of CAD in terms of the most reliable indicator, i.e. absolute detection rates at screening. Such a benefit, if any, may be difficult to prove satisfactorily at centers that perform a low volume of the examination.

5. Conclusions

The utilization of the R2 ImageChecker at a low volume mammography center resulted in an apparent small (5.0%) increase in the detection of impalpable breast cancer while an additional 1.0% of patients were recalled for additional imaging as a result of CAD prompts. The benefits of CAD at clinics doing a low volume of mammography may be less than those obtained at high-volume centers.

Acknowledgements

This work was supported in part by U.S. Army DAMD 17-96-1-6228. The authors thank Timothy Merrill, M.D., Joseph Chessare, M.D. and Allan Haggar, M.D. for their participation in this study. Melissa Tischman, R.T.(R)(M) and Geraldine Calzaretta, R.T.(R)(M) performed the mammograms during the period described in this paper. C. Vyborny and K. Doi are shareholders in R2 Technology Inc.

References

1. Nishikawa, R.M., Giger, M.L., Schmidt, R.A., Wolverton, D.E., Doi K.: Prospective Testing of a Clinical CAD Workstation for the Detection of Breast Lesions on Mammograms. In: Doi, K., MacMahon, H., Giger, M.L., Hoffmann K.R. (eds.) Computer-Aided Diagnosis in Medical Imaging: Proceedings of the 1st International Workshop on Computer-Aided Diagnosis, Elsevier, Amsterdam (1997) 209-214
2. Burhenne, L.J.W., D'Orsi, C.J., Feig, S.A., Kopans, D.B., Sickles, E.A., Tabar, L., Vyborny, C.J., Castellino, R.A.: The Potential Contribution of Computer-aided Detection to the Sensitivity of Screening Mammography. Radiology 215 (2000) 554-562
3. Freer, T.W., Ulissey, M.J.: Screening Mammography with Computer-aided Detection: Prospective Study of 12,860 Patients in a Community Breast Center. Radiology 220 (2001) 781-786
4. Cupples, T.E.: Impact of Computer-aided Detection in a Regional Screening Mammography Program. Radiology 221(P) (2001) 520
5. Tonita, J.M., Hillis, J.P., Lim, C.H.: Medical Radiologic Technologist Review: Effects on a Population-based Breast Cancer Screening Program. Radiology 211 (1999) 529-533
6. Wolverton, D.E., Sickles, E.A.: Clinical Outcome of Doubtful Mammographic Findings Identified Prospectively on Screening Mammograms. AJR 167 (1996) 1041-1045

Improvement in the automatic detection of individual microcalcifications to integrate a cluster-detection and a cluster-classification schemes

Maria F. Salfity[a], Robert M. Nishikawa[b],Yulei Jiang[b] and John Papaioannou[b]

[a]Instituto de Física Rosario, CONICET-UNR
Bv. 27 de Febrero 210 bis, 2000 Rosario, Argentina
salfity@ifir.edu.ar
[b]Kurt Rossmann Laboratories for Radiologic Image Research, Department of Radiology, MC2026, The University of Chicago, 5841 South Maryland Avenue, Chicago, IL 60637

Abstract. Computer-aided diagnosis schemes for the detection and classification of microcalcification clusters have been developed at our laboratory. The classification scheme takes as input the location of each microcalcification present in the cluster. The cluster-detection scheme is optimised for the detection of clusters and identifies in average half of the actual calcifications present in each cluster. Therefore, the input of the classification scheme was generated by manually identifying the individual calcifications. In this work we present a calcification-detection scheme that acts as an interface between the cluster-detection and classification programs in order to avoid manual identification of each calcification. The new scheme analyses the region that contains a cluster previously identified by a radiologist or by an initial computer scheme, and requires as input the cluster location and size and the approximate number of calcifications in the cluster. This a priori information is used to identify the individual calcifications more accurately.

1. Mammogram Databases

We used two independent mammogram databases. Database I contained 107 microcalcification clusters (40 malignant, 67 benign) and Database II contained 246 microcalcification clusters (123 malignant, 123 benign). All films were digitized with a grey-scale resolution of 10 bits and a pixel size of 0.1 mm/pixel. A researcher manually identified the microcalcifications present in each film. A region of interest defined by the cluster bounding box plus a 55-pixel margin surrounding it was extracted for each cluster. We used Database I to determine the appropriate parameters of the calcification-detection scheme, and Database II for evaluation.

2. Calcification-detection Scheme

The scheme takes as input a region of interest that contains a cluster of microcalcifications and the class to which the cluster belongs according to its number of calcifications, N. Three classes were used [1]: class 1 if $N < 6$, class 2 if $6 \leq N \leq 10$, and class 3 if $N > 10$.

2.1 Preprocessing

The calcifications were enhanced with a Difference of Gaussians (DoG) filter. The Gaussian kernels of the DoG filter had standard deviation values of 1.1 and 1.4 and kernel sizes of 7x7 and 9x9 pixels respectively [2].

2.2 Segmentation Using a Priori Information

The enhanced potential calcifications were segmented by global and local iterative thresholding operations that used a priori information about the approximate number of calcifications in the cluster. First, a global threshold based on the histogram of the image segmented an initially large number of candidate calcifications. The value of the global threshold was iteratively changed until a certain number of potential calcifications, which was previously defined as a function of the cluster class, were segmented. Second, a local thresholding reduced false-positive detections and maintained a minimum number of potential calcifications, which also depended on the cluster class. The local threshold value was calculated in regions of 100x100 pixels, centred on each previously identified potential calcification, and its value was set as $\mu + k^*\sigma$, where μ and σ are the mean and standard deviation values of the region, and k is a variable parameter. If necessary, the value of k was iteratively decreased 0.01 until the number of segmented signals was above the predetermined minimum value. The output of the calcification-detection scheme consists of the locations of the detected signals and can be used as input to the classification scheme.

3. Results and Conclusions

The performance of the calcification-detection scheme was significantly higher than the performance of the cluster-detection scheme in the task of identifying individual calcifications (Fig. 1). For $k = 3.2$, 85% of the actual calcifications were identified in Database I (Fig. 1). In the independent Database II, 77 % of the calcifications were detected at the same operating point. When these computer-detected calcifications were input to the cluster classifier [3], the classification performance was comparable to that obtained by using manually-identified calcifications (Table 1). In this way, with a minimal human intervention the cluster-detection and classification schemes can be linked by the calcification-detection scheme. In a completely automated system, i.e. when the output of the cluster-

detection scheme constitutes the input of the calcification-detection scheme, the performance of the cluster classifier is degraded (for Database I, per patient A_z=0.67 and per cluster A_z=0.75) as a consequence of the underestimation of the number of calcifications by the cluster detection scheme. A more sophisticated calcification-

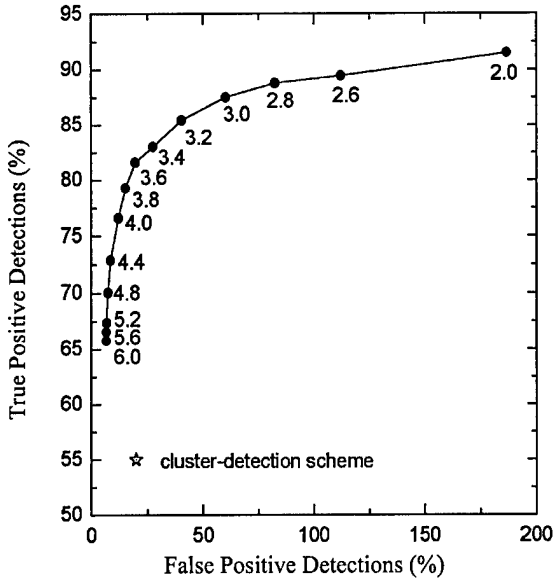


Fig. 1. FROC curve of the calcification-detection scheme for Database I. The different numbers correspond to the value of the local threshold variable parameter *k*. For comparison, the performance of the cluster-detection scheme for the detection of individual calcifications is also indicated on the plot.

Table 1. Area index A_z values obtained by using as input of the cluster-classifier manually-identified calcifications and computer-identified calcifications with the calcification-detection scheme.

Classifier Input	Database I		Database II	
	A_z		A_z	
	Per patient	Per cluster	Per patient	Per cluster
Manual	0.92	0.82	0.91	0.93
Calc-detect. Scheme	0.92	0.83	0.85	0.89

References

1. American College of Radiology: Breast Imaging Reporting and Data System. American College of Radiology, Reston VA (1998)

2. Nishikawa, R.M., Salfity, M.F., Jiang, Y., Papaioannou, J.: Improving the automated classification of clustered calcifications on mammograms through the improved detection of individual calcifications. Proc. SPIE 2002 Medical Imaging (2002) In press
3. Jiang, Y., Nishikawa, R.M., Wolverton, D.E., Metz, C.E., Giger, M.L., Schmidt, R.A., Vyborny, C.J., Doi, K.: Malignant and benign clustered microcalcifications: automated feature analysis and classification. Radiology 198 (1996) 671-678

Effect of radiologists' variability on the performance of computer classification of malignant and benign calcifications in mammograms

Yulei Jiang, M. Fernanda Salfity, Vicky Chen, Robert M. Nishikawa, John Papaioannou,
Alexandra V. Edwards, Sophie Paquerault,
Department of Radiology, The University of Chicago, Chicago, IL 60637

ABSTRACT

In developing a computer technique to classify clustered microcalcifications as malignant or benign, we previously indicated manually the location of all individual calcifications to the computer and found the computer to be more accurate than radiologists. In this study, we investigate whether radiologists can be asked to provide minimal input to the computer and obtain consistent computer classification results. Radiologists were instructed to draw a rectangle that enclosed all calcifications, and indicate the approximate number of the calcifications (either <6, 6–10, 10–30, or >30). The computer used these two pieces of information to detect the individual calcifications and, subsequently, to classify the calcifications as malignant or benign based on only those calcifications detected by the computer. We showed at the 2002 RSNA conference 18 cases together with standard and magnification view mammograms to 38 self-reported breast-imaging radiologists (12 of whom read all 18 cases). The standard deviation in the location of their rectangles (averaged over all cases) was approximately 3 mm, the standard deviation in the linear dimension of the rectangles was 6 mm, and the standard deviation in the computer-estimated likelihood of malignancy was 17%. These results indicate that radiologists are able to provide consistent input to the computer, which in turn produces reasonably consistent computer classification results.

Keywords: computer-aided diagnosis, classification, clustered microcalcifications, reader variability

1. INTRODUCTION

One of the challenges for mammography is that a large number of biopsies are performed on benign lesions because radiologists are not able to differentiate them from malignant lesions. Previous research demonstrates that computer-aided diagnosis holds the potential to help radiologists reduce the number of biopsies of benign lesions while maintaining or even increasing the correct diagnosis of malignant lesions [1-4]. Previous work has concentrated on automated computer analysis of breast lesions. However, in the setting of a diagnostic examination, the radiologist knows exactly where the lesion is in the mammogram. The radiologist may want to query the computer after providing the exact location of the lesion to the computer. The purpose of this study was to investigate radiologists' variability in locating the lesions and any dependence in the results of the computer analysis on such variability.

2. MATERIALS AND METHODS

2.1. Computer classification of calcifications as malignant or benign

The objective of our technique was to use a computer to analyze calcifications in mammograms automatically, to classify them as malignant or benign, and subsequently to make the results of this computer analysis available to radiologists as an aid to their diagnostic decision-making. The technique has been described in detail elsewhere [1, 2, 5]. Briefly, the computer extracts eight image features from digitized mammograms that describe the size and shape of the calcification cluster, the average and variation in size (including contrast) of the individual calcifications, and the degree of shape-irregularity of the individual calcifications. The computer then uses an artificial neural network to merge these

Medical Imaging 2003: Image Perception, Observer Performance, and Technology Assessment,
Dev P. Chakraborty, Elizabeth A. Krupinski, Editors, Proceedings of SPIE Vol. 5034 (2003)

image features into a single output, and subsequently converts this output to an estimate of the likelihood of malignancy. A prerequisite of this technique is the knowledge of the locations of every individual calcification, from which the image features are to be calculated.

2.2. Computer detection of individual calcifications

We recently developed a computer technique to detect individual calcifications based on certain *a priori* information provided by the radiologist, for the purpose of facilitating subsequent computer classification of the calcifications as malignant or benign without requiring manual identification of the individual calcifications [6, 7]. The *a priori* information consisted of a region of interest (ROI) that encompasses the calcifications and an approximate number of the calcifications in four categories: <6, 6-10, 10-30, and >30. This approach may be appropriate in the settings of diagnostic mammography in which the radiologist knows exactly where the calcifications are, and can quickly count their number. Note that this technique, which may be referred to as the "calcification detection technique," is different from other computer detection techniques that may be referred to as "cluster detection techniques" in that the purpose of this technique is to detect the individual calcifications, rather than to detect the lesion or the group (cluster) of calcifications. We have shown previously that use of this technique can produce comparable computer classification performance of malignant and benign calcifications as manual identification of the calcifications, even though the computer does not detect all calcifications, and does include false-positive signals.

With the use of the "calcification detection technique," the computer classification of malignant and benign calcifications and the computer-estimated likelihood of malignancy become potentially dependent on the *a priori* information provided by the radiologist. The ROI defines the area in a mammogram from which the computer detects calcifications. If, on one hand, the ROI is too small, then computer classification will be based on only some of the calcifications in a cluster. If, on the other hand, the ROI is too large, then computer classification will be based on a collection of "calcifications" that may include many false-positive signals. Similarly, the number of calcifications defines how many calcifications the computer will detect and on which computer classification will be based. If the number is too small, then, again, computer classification will be based on only some of the calcifications in a cluster. If the number is too large, then computer classification will be based on a collection of "calcifications" that include many false-positive signals. Therefore, the extent to which radiologists can provide the *a priori* information accurately and consistently, and the extent to which computer classification is dependent on the *a priori* information need to be evaluated.

2.3. A computer-user interface

For the purpose of this study, a computer-user interface was designed, and is shown in Fig. 1. The purpose of this computer-user interface was to facilitate a radiologist to provide the *a priori* information to the computer, watch the computer perform its analysis on-line (detecting calcifications and classifying the calcifications as malignant or benign), and review the results of the computer calculation. The information shown in Fig. 1 appears in stages, rather than all at once as shown in the figure. Initially, only the mammogram is shown, and the radiologist is asked to draw a rectangular box that is large enough to enclose all suspect calcifications that need to be targeted in a stereo core biopsy but not large enough to indicate a broad general area. The radiologist is also asked to indicate an approximate number of calcifications and is warned that computer calculation may be sensitive to the calcification numbers indicated. Subsequent to computer calculations, the radiologist is given opportunities to alter the ROI and the calcification number count and have the computer repeat its calculations. After entering the *a priori* information, the radiologist is asked to enter BI-RADS assessment in terms of the need for biopsy before and after reviewing the results of computer calculations. Results of the computer calculations are shown in two forms. An annotated ROI is shown in the lower-right corner with black dots indicating the "calcifications" detected by the computer (which might also include false-positive signals). The radiologist could use this ROI to assess whether the computer has detected all calcifications and whether the computer detection has included false-positive signals. The computer-estimated likelihood of malignancy is shown in the upper right corner. Results are shown separately for the two views of the mammogram. At the conclusion of each case, the radiologist reviews the histological diagnosis. In addition and not shown in Fig. 1, the final display also

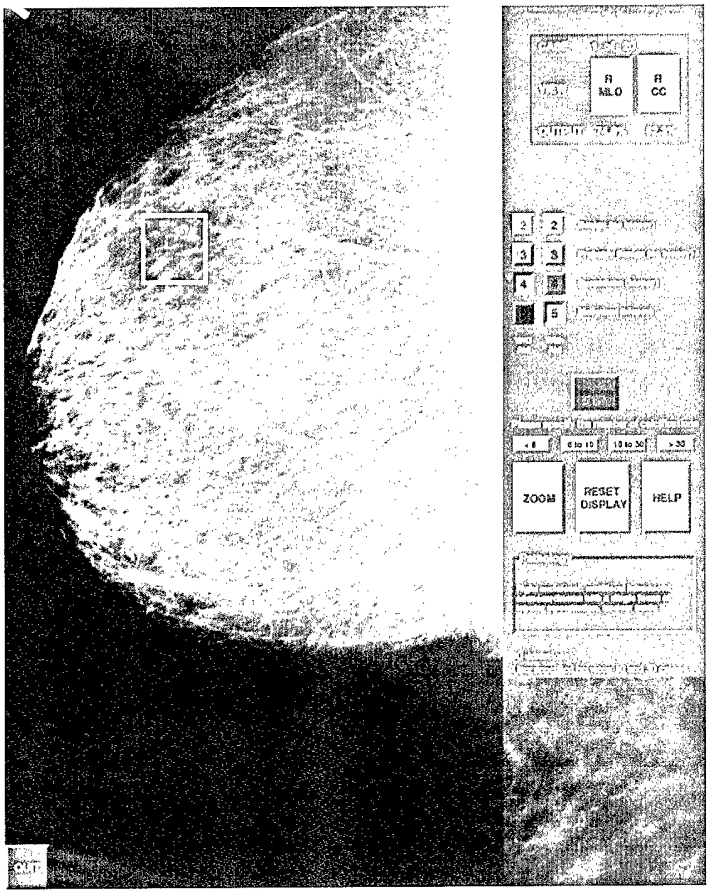


Figure 1. An example of the computer-user interface for the radiologist to provide *a priori* information to the computer and to review results of the computer calculations. Information is displayed in stages, rather than all at once. See text for description of the sequence of display.

includes information on the ROIs drawn by other radiologists, standard deviations in the computer-estimated likelihood of malignancy, and standard deviations in the BI-RADS assessments, if that information is available.

2.4. Study design and data collection

We conducted an experiment as part of an Educational Exhibit at the 88th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, 1-6 December 2002 [8]. Using the interface described above, we showed 18 cases of mammograms that consisted of a lateral view and a CC view containing calcifications. To provide sufficient diagnostic information, we also provided high-quality copy films mounted on a film viewer of the standard views of both breasts and magnification views of the calcifications. The readers were instructed to review the films. During the conference, 38 self-described radiologists who conduct breast imaging more than 10% of their clinical practice read at least some cases, and 12 radiologists read all 18 cases. We recorded the ROIs and the estimated calcification numbers provided by the radiologists, and the results of computer calculations based on the radiologists' input.

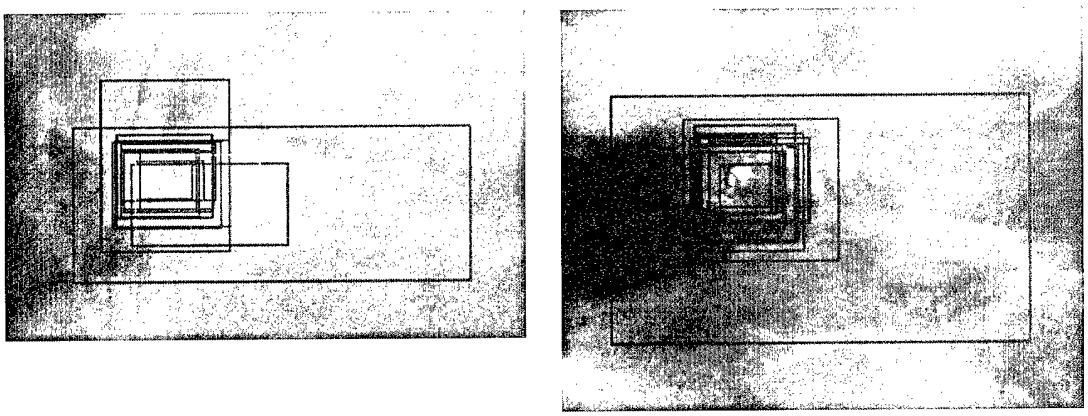


Figure 2. An example of the ROIs drawn by 13 radiologists in the MLO (left) and CC (right) view mammograms of a case diagnosed as fibroadenoma with coarse stromal calcifications, fibrocystic changes.

3. RESULTS

Radiologists were consistent in providing the ROIs and in estimating the number of calcifications. The standard deviation in the center location of the ROIs was 3.0 mm averaged over the horizontal and vertical directions. This was averaged over the two images in each case, and averaged over all cases. The standard deviation in the width and height of the ROIs averaged 5.8 mm, or 40% of the width or height. In no case did all the radiologists agree on one number category for the calcifications in each image. In most cases, the radiologists' input spanned two number categories. In eight cases, the radiologists' input spanned three categories, but only 1-3 radiologists selected the third number category—therefore, this was uncommon. In only one case—in which the calcifications were extremely difficult to see even on the film—did the radiologists select all four number categories.

The computer calculation was influenced by the variability in the *a priori* information provided by the radiologists. The average standard deviation in the number of detected calcifications was 10, and this was averaged over all images and over all cases. The average relative standard deviation in the number of detected calcifications was 38%. The average standard deviation in the computer-calculated likelihood of malignancy was 17%, over all images and over all cases.

Figure 2 shows an example case of the ROIs drawn by 13 radiologists. The average standard deviation in the center position of the ROIs was 1.3 mm and the average standard deviation in the width and height of the ROIs was 4.7 mm. Note that much of the variability is contributed from 1 or 2 ROIs that are appreciably larger than others. Similar observation is made in other cases, except that it tended not to be the same radiologists who drew larger ROIs in each case. The computer-estimated likelihood of malignancy was 13±5% for the MLO view and 10±4% for the CC view. Therefore, the computer calculation was not very sensitive to the variability in the size of the ROIs, or at least less so than it was sensitive to variability in the calcification number categories selected by the radiologists. Figure 3 shows another example image of the ROIs drawn by 38 radiologists. The average standard deviation in the center location of the ROIs was 1.4 mm and the average standard deviation in the width and height of the ROIs was 5.3 mm. These values are not dissimilar to the corresponding values in the previous case. However, the computer-estimated likelihood of malignancy was 67±22%. The larger variation in the computer calculations was not caused by the two largest ROIs. Rather, it was caused by inputs from 3 radiologists who indicated that the number of calcifications was 6-10 in this

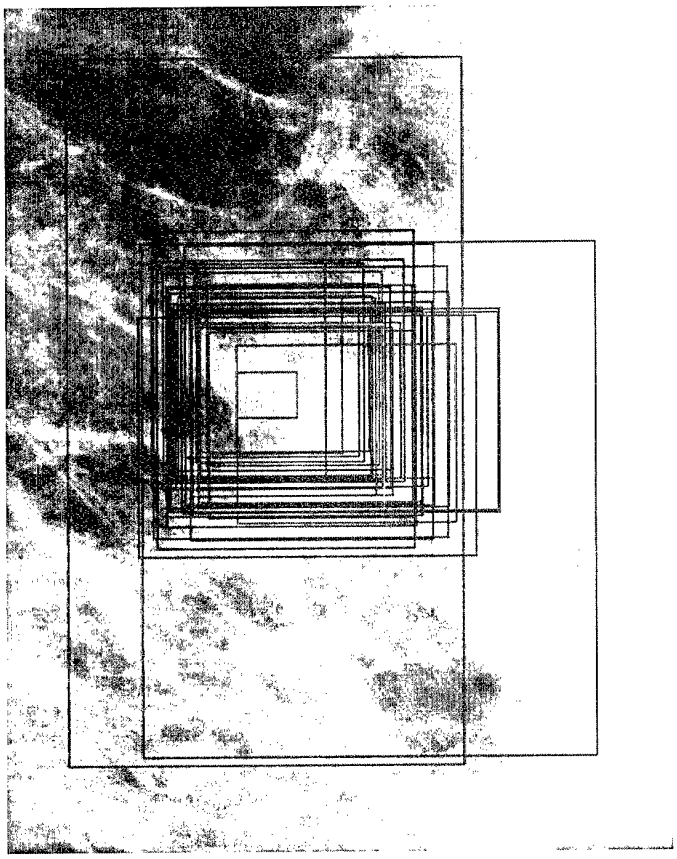


Figure 3. The MLO view of a case diagnosed as DCIS, cribriform and solid type, low to high grade, with focal comedo necrosis and calcifications.

image, and by input from another radiologist who drew the smallest ROI in the center. If these four inputs were eliminated, then the computer-estimated likelihood of malignancy would have been 74±7%.

4. DISCUSSION AND SUMMARY

The task of selecting an ROI and estimating the approximate number of calcifications is not a task that radiologists routinely perform in diagnostic examinations. The results from our experiment indicate that radiologists are able to select ROIs in a consistent way, with only minimal written instructions. This, in part, is because in their clinical practice, radiologists need to estimate the spatial extent of the calcifications, which is akin to selecting an ROI. However, some radiologists have the tendency to select overly large ROIs, perhaps with the thinking of "being on the safe side," or perhaps not realizing the need to be precise in selecting the ROIs. The consistency in selecting the ROIs can possibly increase with more clearly defined instructions. Our results also indicate that radiologists were reasonably consistent in estimating the number of calcifications. However, the computer calculation was more sensitive to variability in the calcification number categories than it was to variability in the ROIs. Radiologists do not routinely count the precise number of calcifications. In our study, they appeared not to appreciate the importance of this estimate for the computer-calculated likelihood of malignancy. Therefore, a danger exists that a radiologist might estimate the

number of calcifications haphazardly, and subsequently misconstrue an incorrect computer-calculated likelihood of malignancy that was based on the incorrect estimate of the number of calcifications. The solution to this problem is to device a way to eliminate the dependence of the computer-calculated likelihood of malignancy on the user-estimated number of calcifications. Despite various sources of variability in the input provided by radiologists and despite any possible exaggeration of the variability because of the RSNA conference environment in which we conducted the experiment, the computer-calculated likelihood of malignancy was reasonably consistent. Therefore, we are optimistic that this approach of using a computer to detect calcifications and to classify them as malignant or benign can eventually become a useful clinical tool to radiologists.

5. ACKNOWLEDGEMENTS

This work was supported in part by NIH/NCI (CA60187, CA092361), the US Army Medical Research and Materiel Command (DAMD17-00-1-0197), and the Susan G. Komen Breast Cancer Foundation. RM Nishikawa and J Papaioannou are shareholders of R2 Technology Inc., Sunnyvale, CA.

6. REFERENCES

[1] Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi, "Improving breast cancer diagnosis with computer-aided diagnosis," *Acad Radiol*, vol. 6, pp. 22-33, 1999.

[2] Y. Jiang, R. M. Nishikawa, R. A. Schmidt, A. Y. Toledano, and K. Doi, "The potential of computer-aided diagnosis (CAD) to reduce variability in radiologists' interpretation of mammograms depicting microcalcifications," *Radiology*, vol. 220, pp. 787-794, 2001.

[3] H. P. Chan, B. Sahiner, M. A. Helvie, N. Petrick, M. A. Roubidoux, T. E. Wilson, D. D. Adler, C. Paramagul, J. S. Newman, and S. Sanjay-Gopal, "Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study," *Radiology*, vol. 212, pp. 817-827, 1999.

[4] Z. Huo, M. L. Giger, C. J. Vyborny, and C. E. Metz, "Breast cancer: Effectiveness of computer-aided diagnosis -- Observer study with independent database of mammograms," *Radiology*, vol. 224, pp. 560-568, 2002.

[5] Y. Jiang, R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vyborny, and K. Doi, "Malignant and benign clustered microcalcifications: automated feature analysis and classification," *Radiology*, vol. 198, pp. 671-678, 1996.

[6] R. M. Nishikawa, M. F. Salfity, Y. Jiang, and J. Papaioannou, "Improving the automated classification of clustered calcifications on mammograms through the improved detection of individual calcifications," *Proc. SPIE*, vol. 4684, pp. 1339-1345, 2002.

[7] M. F. Salfity, R. M. Nishikawa, Y. Jiang, and J. Papaioannou, "The use of a priori information in the detection of mammographic microcalcifications to improve their classification," *Med Phys*, (in press).

[8] Y. Jiang, R. M. Nishikawa, M. L. Giger, J. Papaioannou, L. Lan, and C. J. Vyborny, "On-line demonstration of computer-aided diagnosis (CAD) of malignant and benign breast lesions," *Radiology*, vol. 225(P), pp. 683, 2002.

Automated selection of BI-RADS lesion descriptors for reporting calcifications in mammograms

Sophie Paquerault[1a], Yulei Jiang[a], Robert M. Nishikawa[a], Robert A. Schmidt[a], Carl J. D'Orsi[b], Carl J. Vyborny[a], Gillian M. Newstead[a]

[a]Dept. of Radiology, Univ. of Chicago, 5841 S. Maryland Ave., MC2026, Chicago, IL 60637;
[b]Breast Imaging Center, Emory University, 1365-B Clifton Rd., NE, Atlanta, GA 30322

ABSTRACT

We are developing an automated computer technique to describe calcifications in mammograms according to the BI-RADS lexicon. We evaluated this technique by its agreement with radiologists' description of the same lesions. Three expert mammographers reviewed our database of 90 cases of digitized mammograms containing clustered microcalcifications and described the calcifications according to BI-RADS. In our study, the radiologists used only 4 of the 5 calcification distribution descriptors and 5 of the 14 calcification morphology descriptors contained in BI-RADS. Our computer technique was therefore designed specifically for these 4 calcification distribution descriptors and 5 calcification morphology descriptors. For calcification distribution, 4 linear discriminant analysis (LDA) classifiers were developed using 5 computer-extracted features to produce scores of how well each descriptor describes a cluster. Similarly, for calcification morphology, 5 LDAs were designed using 10 computer-extracted features. We trained the LDAs using only the BI-RADS data reported by the first radiologist and compared the computer output to the descriptor data reported by all 3 radiologists (for the first radiologist, the leave-one-out method was used). The computer output consisted of the best calcification distribution descriptor and the best 2 calcification morphology descriptors. The results of the comparison with the data from each radiologist, respectively, were: for calcification distribution, percent agreement, 74%, 66%, and 73%, kappa value, 0.44, 0.36, and 0.46; for calcification morphology, percent agreement, 83%, 77%, and 57%, kappa value, 0.78, 0.70, and 0.44. These results indicate that the proposed computer technique can select BI-RADS descriptors in good agreement with radiologists.

Keywords: Mammography, computer-aided diagnosis, BI-RADS, microcalcifications, classification, linear discriminant analysis.

1. INTRODUCTION

Mammography is the only proven screening technique for detecting breast cancer in its early stages [1]. However, the analysis and interpretation of mammograms for the diagnosis of breast cancer are difficult tasks. When finding abnormal lesions on mammograms, radiologists often call the patient back for further diagnostic evaluations and frequently recommend a biopsy to avoid missing any breast cancer. Computer-aided diagnosis (CAD) systems have been developed to provide a second opinion to radiologists. These systems use computer vision and pattern recognition techniques to automatically detect and characterize abnormal lesions on mammograms. Although it has been reported that these systems are useful in reducing the error rate in mammographic screening [2, 3] and in lowering the biopsy recommendation rate for benign breast lesions [4-6], the detection sensitivity and differentiation of malignant from benign lesions need to be improved to provide maximum benefit to the radiologist and the patient.

To further improve the accuracy in mammography reporting, the American College of Radiology designed a standardized lexicon: the Breast Imaging Reporting and Data System (BI-RADS) [7]. A few studies have demonstrated the potential of using this standard lexicon for both radiologists and CAD systems [8-10]. Hara et al. [11] proposed a technique to automatically determine the microcalcification distribution as described in BI-RADS and incorporated this

[1] paquerau@uchicago.edu; phone 1 773 834-5094 ; fax 1 773 702-0371

Medical Imaging 2003: Image Processing, Milan Sonka, J. Michael Fitzpatrick, Editors,
Proceedings of SPIE Vol. 5032 (2003) © 2003 SPIE · 1605-7422/03/$15.00

information into their current mammographic microcalcification classification technique. Although the number of cases tested was small, they demonstrated a significant improvement of the microcalcification classification performance. However, the entire BI-RADS information has not yet been utilized in the CAD systems.

For clustered microcalcifications, BI-RADS contains 5 calcification distribution descriptors and 14 calcification morphology descriptors. Incorporating the output of a computer system that automatically provides the calcification description according to the BI-RADS lexicon to the current CAD scheme can possibly improve its accuracy for the differentiation of benign from malignant clustered microcalcifications. We thus propose in this study a method that automatically outputs the calcification distribution descriptor and the calcification morphology descriptor as defined in BI-RADS.

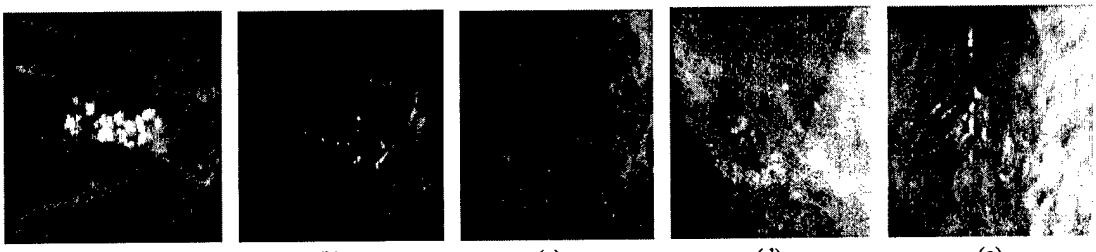
2. MATERIALS AND METHODS

In this study, as for any pattern recognition problem, we first extracted features of the segmented calcifications. Two subsets of these features were retained that characterized at least one of the calcification distribution descriptors or at least one of the calcification morphology descriptors. Following this step, a classification scheme was designed that used several linear discriminant analyses (LDAs). The computer output consisted of one calcification distribution descriptor and two calcification morphology descriptors. The results were further compared to the radiologists' data in terms of percent agreement and kappa values.

2.1 Database

The database consisted of 90 cases of mammograms containing clustered microcalcifications. The mammograms were selected from patient files in the Department of Radiology at the University of Chicago. Ninety two per cent of them were composed of the two standard craniocaudal (CC) and medio-lateral oblique (MLO) views. These mammograms were digitized with a LUMISYS laser film scanner at a pixel size of 100 μm and 12-bit gray levels. A total of 172 images were available for our study. The locations of the microcalcifications were identified manually to avoid incorporating false positives or missing any true calcifications in the subsequent analysis.

Three expert radiologists reviewed our database and described the calcifications using the BI-RADS descriptors. The radiologists were allowed to report one calcification distribution descriptor and up to two calcification morphology descriptors because it was often difficult to describe calcification morphology with a single descriptor. In our study, the radiologists used only 4 of the 5 calcification distribution descriptors (grouped, linear, segmental and regional), and 5 of the 14 calcification morphology descriptors contained in BI-RADS (punctate, amorphous, pleomorphic, fine linear branching, and coarse). The radiologist occasionally used the terms dystrophic and round. But because these were not apparently used in a consistent way and because the number of cases for each of these descriptors was small, we combined these cases with those reported as coarse. Figure 1 shows examples of the 5 calcification morphology descriptors reported by one radiologist. Histograms of the calcification distribution descriptors and calcification morphology descriptors rated by each radiologist on this dataset are shown in Figs. 2 and 3. The BI-RADS data reported by one of these three radiologists was used to train our method, and subsequently the BI-RADS descriptors reported by all three radiologists were compared to the computer output.



(a) (b) (c) (d) (e)

Fig. 1: Example images of the 5 calcification morphology descriptors reported by the radiologists (a) coarse, round and dystrophic, (b) punctate, (c) amorphous, (d) pleomorphic, and (e) fine linear branching calcifications.

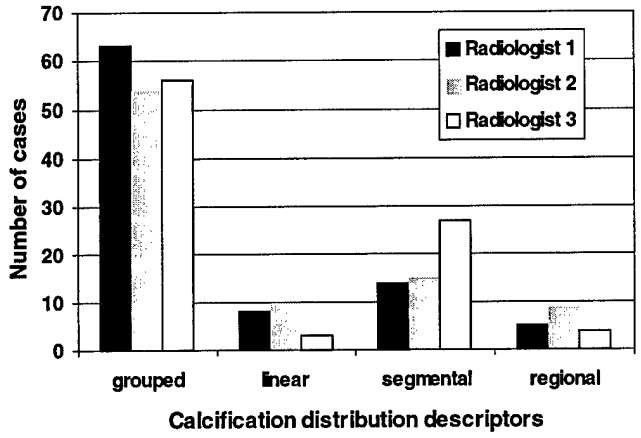


Fig. 2: Histograms of the calcification distribution descriptor rated by 3 expert mammographers for a database of 90 cases of digitized mammograms.

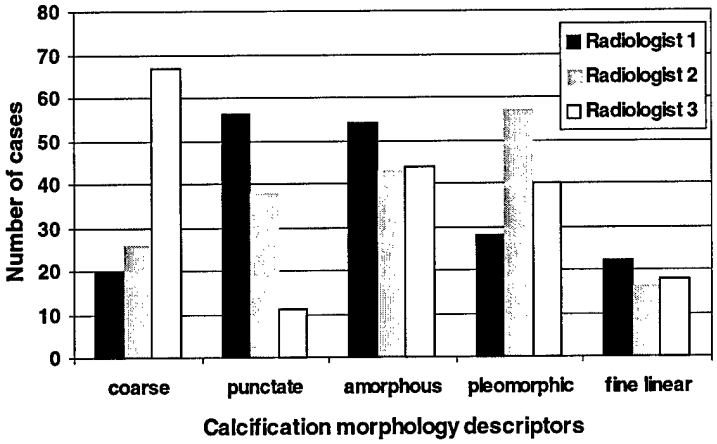


Fig. 3: Histograms of the calcification morphology descriptor rated by 3 expert mammographers for a database of 90 cases of digitized mammograms. The radiologists often reported two different BI-RADS descriptors per cluster; these two descriptors were included in these histograms.

2.2 Methods

A segmentation stage was first applied to extract microcalcifications on each mammogram. This stage was detailed in [12]. To avoid incorporating false positives and missing any calcifications during this segmentation stage, the locations of the calcifications on each mammogram were manually identified. Features based on the size and shape of the segmented calcifications, and the size and shape of the cluster were then extracted for further analysis [13]. A total of 39 features were defined, but because of the limited number of cases in our data set, no automatic feature selection methods were used [14]. Based on the BI-RADS definitions and an intuitive understanding of the calcification distribution and morphology descriptors, we manually selected two subsets of features. The performance of these two sets of features is reported.

These two subsets of features were then used as input respectively to two separate sets of classifiers. The first set of classifiers was designed to select the most appropriate calcification distribution descriptor, and the second set of classifiers was for the calcification morphology descriptor. For the calcification distribution descriptor, 4 linear discriminant analysis classifiers were developed using the first subset of features based on the size and shape of the cluster, and were trained using the BI-RADS data reported by the first radiologist. Each LDA evaluated one of the BI-RADS calcification distribution descriptors against the others. For example, a LDA was defined to characterize regional calcification clusters from all other calcification distribution descriptors. To evaluate our method in an unbiased way with the data from the first radiologist, we used the leave-one-out technique, which consists of training on all cases minus one, and testing on the remaining case. These classifiers produced 4 calcification distribution scores (group 1) of how well each descriptor describes a cluster. A similar technique was used for the calcification morphology descriptors. Five LDAs were designed using the second subset of features based on the size and shape of the individual calcifications. These LDAs were also trained using the calcification morphology descriptors reported by the first radiologist. Because radiologists were allowed to report up to two calcification morphology descriptors, only the first reported descriptor, considered to be the most significant, was used to train the LDAs. Five calcification morphology scores (group 2) were produced by these LDAs. Receiver operating characteristic (ROC) analysis was then applied to evaluate the performance of each LDA, and areas under the ROC curves (A_z) are reported.

A final decision was made to determine the computer-identified BI-RADS descriptors. For calcification distribution, we selected the descriptor that corresponded to the maximum score from group 1 and from the two images in each case. For calcification morphology, two descriptors were determined and corresponded to the two highest scores from group 2 and from the two images in each case. We then compared the computer-selected descriptors to the BI-RADS descriptors provided by each radiologist, in terms of percent agreement and the unweighted kappa. These measures of agreement were also computed for pairs of the 3 radiologists.

3. RESULTS AND DISCUSSION

The first stage of our proposed method consisted of the selection of two subsets of discriminant features. Based on the definition of the calcification distribution descriptors, 5 computer-extracted features were selected and consisted of the area of the microcalcification cluster, the perimeter, the circularity, and the mean and standard deviation of the distance between pairs of calcifications in the cluster. Figure 4 shows the performance of these features on an individual basis to discriminate each calcification distribution descriptor against all others that are available. Based on the definition of the calcification morphology descriptors, 10 computer-extracted features were selected and these consisted of the relative standard deviation of the calcification width, the mean thickness, the mean volume, the fraction of calcifications with centroid outside the segmented calcification, the mean of the standard deviation of the shape index, the mean, the standard deviation, the first, and the second highest value of the relative standard deviation of the shape index, and the maximum of the ratio of the calcification perimeter to its area. Figure 5 shows the performances of these features on an individual basis to discriminate each calcification morphology descriptor against all others that are available. We constrained the study to these two subsets of features but will investigate other relevant features in the future in order to classify other calcification distribution and morphology descriptors that were not present in sufficient number in our database.

These two subsets of selected features were used as input to a set of 4 and a set of 5 different LDAs to select the calcification distribution and morphology descriptors. Tables 1 and 2 show the performance of each LDA on the training and test sets. For calcification distribution, the highest performance was obtained for identifying the segmental and grouped descriptors. This could be related to the large number of cases contained in our dataset for these two descriptors. For calcification morphology, the highest performance was obtained for the classification of coarse calcification cases. This could be explained by the relative obvious difference in size and shape of this type of calcifications compared to the other ones. Further tests on an independent database will be required to test the validity of these results.

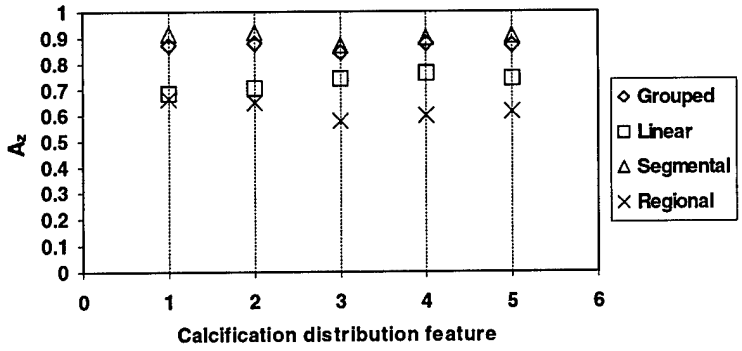


Fig. 4: A_z values of each individual feature in being used as a basis to identify one calcification distribution descriptor as the most appropriate for a given cluster. The features are: (1) the area of the microcalcification cluster, (2) the perimeter, (3) the circularity, (4) the mean and (5) standard deviation of the distance between pairs of calcifications in the cluster.

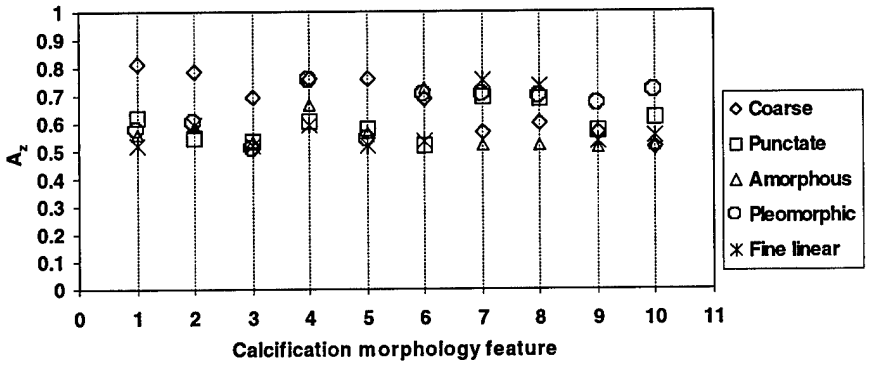


Fig. 5: A_z values of each individual feature in being used as a basis to identify one calcification morphology descriptor as the most appropriate for a given cluster. The features are: (1) relative standard deviation of the calcification width, (2) mean thickness, (3) mean volume, (4) fraction of calcifications with centroid outside the segmented signal, (5) mean of the standard deviation of the shape index, (6) the mean, (7) standard deviation, (8) first and (9) second highest value of the relative standard deviation of the shape index, and (10) maximum of the calcification perimeter to area ratio.

TABLE 1: Performance of the 4 calcification distribution LDAs on the training and test sets.

	Grouped	Linear	Segmental	Regional
Training set	0.88	0.83	0.93	0.73
Test set	0.85	0.76	0.91	0.64

Table 2: Performance of the 5 calcification morphology LDAs on the training and test sets.

	Coarse	Punctate	Amorphous	Pleomorphic	Fine linear
Training set	0.92	0.75	0.78	0.80	0.86
Test set	0.82	0.68	0.70	0.71	0.74

We compared the computer-provided BI-RADS descriptors to the BI-RADS data reported by the three radiologists. For calcification distribution, the results of the evaluation are detailed in Table 3 and indicate that the agreement between

computer-provided calcification distribution descriptor and the BI-RADS data provided by the three radiologists were similar. The kappa values, κ=0.44, κ=0.36, and κ=0.46, demonstrate moderate agreement between the computer and each radiologist. For calcification morphology, two evaluations were made. The first evaluation consisted of using the two computer-provided calcification morphology descriptors and comparing them to the radiologists' BI-RADS data. It was considered to be a success when at least one of the two computer descriptors corresponded to one of the two descriptors reported by the radiologist. This evaluation is reported in Table 4 (A) and indicates that the computer output was somewhat close to the radiologists' descriptions of the calcification morphology. The kappa values, κ=0.78, κ=0.70, and κ=0.44, demonstrate that the computer is in good agreement with the first two radiologists, and moderate agreement with the third radiologist. The second evaluation consisted of using the highest computer-provided calcification morphology descriptor and comparing it to the radiologists' data. It was considered to be a success when the computer-provided descriptor corresponded to one of the two descriptors reported by the radiologist. As shown in Table 4 (B), the kappa values, κ=0.45, κ=0.42, and κ=0.15, demonstrate moderate agreement with the first two radiologists, and poor agreement with the third radiologist. This lower agreement means that two calcification morphology descriptors provided by the computer are needed to be consistent with the radiologists' interpretation of the mammograms.

We also compared pairs of the three radiologists' BI-RADS data in comparison to the computer-provided BI-RADS descriptors. For calcification distribution, the resulting evaluation is detailed in Table 5. The kappa values, κ=0.53, κ=0.50, and κ=0.37, demonstrate a moderate agreement. The computer output is thus consiste nt with the radiologists'interpretation of the calcification distribution on mammograms. For calcification morphology, as the radiologists were allowed to report up to two descriptors, we used a similar evaluation as the first evaluation for the computer. It was considered to be a success when at least one of the two descriptors reported by one radiologist corresponded to one of the two descriptors reported by another radiologist. The results are detailed in Table 6. The kappa values, κ=0.73, κ=0.50, and κ=0.64, demonstrate that the three radiologists were in good agreement. Similar agreement was obtained when comparing the computer output to the radiologists' data. This demonstrates that the computer method is also consistent with the radiologists' interpretation of calcification morphology on mammograms.

TABLE 3: Comparison of the computer-selected calcification distribution descriptors to the BI-RADS descriptors reported by the three radiologists.

	Percent agreement	κ value
Radiologist 1	74	0.44
Radiologist 2	66	0.36
Radiologist 3	73	0.46

TABLE 4: Comparison of the computer-selected calcification morphology descriptors to the BI-RADS descriptors reported by the three radiologists. A: It was considered to be a success when at least one of the two computer-identified calcification morphology descriptors corresponded to one of the two calcification morphology descriptors reported by the radiologist. B: It was considered to be a success when the first of two computer-identified calcification morphology descriptors corresponded to one of the two calcification morphology descriptors reported by the radiologist.

	A		B	
	Percent agreement	κ value	Percent agreement	κ value
Radiologist 1	83	0.78	58	0.45
Radiologist 2	77	0.70	54	0.42
Radiologist 3	57	0.44	32	0.15

TABLE 5: Comparison of the calcification distribution descriptors reported by pairs of the three radiologists.

	Percent agreement	κ value
Radiologist 1 / Radiologist 2	75	0.53
Radiologist 1 / Radiologist 3	74	0.50
Radiologist 2 / Radiologist 3	65	0.37

TABLE 6: Comparison of the calcification morphology descriptors reported by pairs of the three radiologists. It was considered to be a success when at least one of the calcification morphology descriptors reported by the radiologists was the same.

	Percent agreement	κ value
Radiologist 1 / Radiologist 2	79	0.73
Radiologist 1 / Radiologist 3	60	0.50
Radiologist 2 / Radiologist 3	72	0.64

4. CONCLUSION

In this study, we propose a method to automatically select the BI-RADS descriptors of clustered microcalcifications. Both the calcification distribution and the calcification morphology descriptors have been analyzed. Based on our understanding of the BI-RADS lexicon, features have been selected and several LDAs have been trained to select the appropriate descriptors. Membership scores are obtained from the different LDAs and a final decision defines the calcification distribution descriptor and the calcification morphology descriptors for each mammogram. The results are compared to the descriptors reported by three radiologists. The preliminary results demonstrate that the proposed method is consistent with the radiologists' interpretation of both calcification distribution and morphology on mammograms. Further tests are underway to extract more relevant features, to integrate and describe calcification distribution and calcification morphology that were not present in our database and to test the robustness of our proposed computer scheme. In addition, we will separately analyze the benefit on the radiologists' interpretation of the mammograms when reading with the proposed computer-selected BI-RADS descriptors, and on the classification of malignant versus benign microcalcification clusters when this proposed computer-selected BI-RADS descriptors are included in the current CAD scheme.

Acknowledgments

This work was supported in part by a grant from Illinois Department of Public Health, a grant from the US Army (DAMD 17-00-1-0197), and a grant from National Institutes of Health (R01 CA092361). The authors are grateful to Charles E. Metz, PhD, for use of the LABROC program.

References

[1] L. Tabar, A. Gad, L.H. Holmberg, U. Ljungquist, G. Eklund, C.J.G. Fagerberg, L. Baldetorp, O. Grontoft, B. Lundstrom, J.C. Manson, N.E. Day, and F. Pettersson, *Reduction in mortality from breast cancer after mass screening with mammography,* Lancet **1** (8433), 829-832 (1985).

[2] H.P. Chan, K. Doi, C.J. Vyborny, R.A. Schmidt, C.E. Metz, K.L. Lam, T. Ogura, Y. Wu, and H. MacMahon, *Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis,* Invest. Radiol. **25**, 1102-1110 (1990).

[3] L.J. Warren Burhenne, S.A. Wood, C.J. D'Orsi, S.A. Feig, D.B. Kopans, K.F. O'Shaughnessy, E.A. Sickles, L. Tabar, C.J. Vyborny, and R.A. Castelino, *Potential contribution of computer-aided detection to the sensitivity of screening mammography,* Radiology **215**, 554-562 (2000).

[4] Y. Jiang, R.M. Nishikawa, R.A. Schmidt, C.E. Metz, M.L. Giger, and K. Doi, *Improving breast cancer diagnosis with computer-aided diagnosis,* Acad. Radiol. **6**, 22-33 (1999).

[5] H.P. Chan, B. Sahiner, M.A. Helvie, N. Petrick, M.A. Roubidoux, T.E. Wilson, D.D. Adler, C. Paramagul, J.S. Newman, and S. Sanjay-Gopal, *Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: An ROC study*, Radiology **212**, 817-827 (1999).

[6] D.J. Getty, R.M. Pickett, C.J. D'Orsi, J.A. Swets, *Enhanced interpretation of diagnostic images*, Invest. Radiol. **23**, 240-252 (1988).

[7] American College of Radiology (ACR), *Breast imaging reporting and data system (BI-RADS™).* Third Edition. American College of Radiology, Reston, VA (1998).

[8] J.A. Baker, P.J. Kornguth, J.Y. Lo, M.E. Williford, and C.E. Floyd, *Breast cancer: Prediction with artificial neural network based on BI-RADS standardized lexicon,* Radiology **196**, 814-822 (1995).

[9] L. Liberman, A.F. Abramson, F.B. Squires, J.R. Glassman, E.A. Morris, and D.D. Dershaw, *The breast imaging reporting and data system: Positive predictive value of mammographic features and final assessment categories,* AJR **171**, 35-40 (1998).

[10] S.G. Orel, N. Kay, C. Reynolds, and D.C. Sullivan, *BI-RADS Categorization as a predictor of malignancy,* Radiology **211**, 845-850 (1999).

[11] T. Hara, A. Yamada, H. Fujita, T. Iwase, and T. Endo, *Automated classification method of mammographic microcalcifications by using artificial neural network and ACR BI-RADS™ criteria for microcalcification distribution,* Proceedings of the 5th International Workshop on Digital Mammography (Medical Physics Publishing, Madison, WI) 198-204 (2000).

[12] Y. Jiang, R.M. Nishikawa, M.L. Giger, K. Doi, R.A. Schmidt, and C.J. Vyborny, *Method of extracting signal area and signal thickness of microcalcifications from digital mammogams,* Proceeding of SPIE **1778**, 28-36 (1992).

[13] Y. Jiang, R.M. Nishikawa, D.E. Wolverton, C.E. Metz, M.L. Giger, R.A. Schmidt, C.J. Vyborny, and K. Doi, *Malignant and benign clustered microcalcifications: Automated feature analysis and classification,* Radiology **198**, 671-678 (1996).

[14] M.A. Kupinski, and M.L. Giger, *Feature selection with limited datasets,* Med. Phys. **26**, 2176-2182, (1999).